Australian Government
Department of Defence
Defence Science and
Technology Organisation

Range Safety Application of Kernel Density Estimation

Gary Glonek[1], Timothy Staniford[2], Michael Rumsewicz[2], Oleg Mazonka[3], Jeremy McMahon[2], Duncan Fletcher and Michael Jokic

Weapons Systems Division
Defence Science and Technology Organisation

[1] School of Mathematical Sciences, University of Adelaide
[2] TRC Mathematical Modelling, University of Adelaide
[3] Davtec IT

DSTO-TR-2292

ABSTRACT

This report describes the kernel density estimation technique and its application to range safety applications. The kernel density estimation technique is shown to be suitable for developing probabilistic risk assessments from ground impact data generated for guided weapon systems via Monte Carlo simulations. An advantage of this technique is that it can be used to predict the probability density function for minimal simulated ground impacts with apparently random distribution. Several techniques have been proposed to ameliorate the identified limitations of the kernel density estimation technique, including a covariant form for two-dimensional data. Analysis of the available simulated guided weapon ground impact data has identified that around six hundred impact points are sufficient for generating a probability distribution.

RELEASE LIMITATION

Approved for public release

Published by

Weapons Systems Division
DSTO Defence Science and Technology Organisation
PO Box 1500
Edinburgh South Australia 5111 Australia

Telephone: (08) 8259 5555
Fax: (08) 8259 6567

AR-014-543
January 2010

APPROVED FOR PUBLIC RELEASE

Range Safety Application of Kernel Density Estimation

Executive Summary

The Range Safety Template Toolkit (RSTT) development project undertaken by Weapons Systems Division of the Defence, Science and Technology Organisation (DSTO) was scoped to develop probabilistic risk hazard analysis capabilities for guided weapon and sounding rocket trials. The Centre for Defence Communications and Information Networking (CDCIN), formerly known as TRC Mathematical Modelling (TRC), at the University of Adelaide was contracted to undertake research and development in support of the RSTT project. To meet the objectives of the RSTT project, DSTO proposed using Monte Carlo simulations of specific vehicles (including likely failure response modes) to generate ground impact data that could be turned into a probability density function. It is this final aspect that was the focus of the research and development discussed in this report.

To support the research and development work, DSTO provided results (including ground impact data) from Monte Carlo simulations of a generic guided weapon system. Consultation between DSTO and CDCIN identified a number of essential activities for this work:

1. Analyse distributions to understand heterogeneous processes.
2. Develop robust estimation methods: Apply EVT (Extreme Value Theory) and other methods to representative distributions for evaluation of statistical method effectiveness.
3. Develop an understanding of "typical" impact distribution data supplied by Weapons Systems Division of DSTO. The data covered a range of missile launch scenarios and failure modes considered by DSTO to be typical of the data that might be generated for actual weapons systems.
4. Investigate techniques suitable for generating approximate probability density functions representing missile impact data.
5. Investigate the convergence properties of those techniques with increasing size of dataset.
6. Investigate the degree of resolution required to provide impact distributions meaningful to use in a Range Safety Template Toolkit.
7. Investigate techniques for generating range safety templates for scenarios for which simulated data are not directly available, more specifically, to investigate the feasibility of using interpolation techniques for approximating a given scenario from other scenarios for which data exists.
8. Examine the impact of alternative distance metrics on the quality of the impact zone interpolation process.

The research and development activities undertaken by by CDCIN and DSTO have:

1. Qualitatively described the features of impact distribution data that may affect subsequent statistical modelling.
2. Defined a technique, specifically, the use of kernel density estimation, for providing a statistical model of a specific missile impact data set which estimates the probability density function of the impact distribution. The solution proposed here is purely data analytic and as such does not allow for the incorporation of any substantive knowledge.
3. Defined a technique for combining kernel density estimates corresponding to different failure modes within a single operational scenario.
4. Defined a technique for incorporating information on missile Maximum Energy Boundaries into the analysis so as to refine the impact zone probability density function.
5. Defined a technique for using the probability density function together with population density information to obtain estimated injury rates for a given scenario.
6. Defined a technique for using the probability density function together with range boundary information to obtain an estimate for a missile leaving a given range.
7. Defined a technique for using the probability density function to determine a conservative, convex safety exclusion zone with given probability of the missile leaving the zone.
8. Defined an approximate technique for defining a conservative exclusion zone derived from probability density functions of different scenarios.
9. Found that KDE resolutions beyond 16 x 16 and 32 x 32 do not provide significantly more accurate information and hence 16 x 16 or 32 x 32 resolutions appear to be suitable for the development of Range Safety Templates.
10. Found that at least 600 observations (impact data points) should be used in generating KDEs for a given scenario.
11. Identified situations in which the Kernel Density Estimation process is not robust, generally when tight clusters of data points occur within the data set. In such cases the bandwidth parameters automatically generated by the process tend to be very small and the KDE generated consequently “erratic”. This report has suggested one method of dynamic bandwidth calculation to improve the PDF for clustered or non-normal ground impact distributions.
12. Described a covariant form the Kernel Density Estimator for two-dimensional data that robustly predicts the ground impact probability function.
13. Outlined a numerical approach to ensure computationally accurate and efficient results are obtained when using the kernel density estimate technique with real impact data.

The results obtained from the work outlined in this document are essential for the operation of the Range Safety Template Toolkit. RSTT is a capability for the generation of probabilistic risk hazard analyses and weapon danger areas for guided weapon and sounding rocket trials. Due to the large flight ranges of these systems and limited range space for trials, RSTT and its supporting research are important for ensuring that future system trials can be practically conducted in Australia. Importantly, the probabilistic methodologies presented here can potentially be applied to a broad range of applications that require risk hazard analysis including: ballistic munition testing, aircraft flight, orbital re-entry, rocket launches and explosive testing.

Authors

Gary Glonek

School of Mathematical Sciences, University of Adelaide

Gary Glonek is an Associate Professor and Head of the Statistics Discipline at the University of Adelaide and has held various offices within the Statistical Society of Australia, including President of the SA Branch. Gary has undertaken numerous statistical consultancies across a wide range of commercial and scientific areas to clients, in all authoring more than 60 consultancy reports. He worked as an Assistant Professor in Statistics at the University of Chicago from 1987-1989 and as a Lecturer and then Senior Lecturer in Statistics at Flinders University before joining the University of Adelaide in 2000. Gary undertook his undergraduate and postgraduate training at the Flinders University of South Australia, graduating with a PhD in Statistics in 1988. He has written several articles in refereed journals and presented papers at a number of conferences. He has held various research grants and currently holds an Australian Research Council Discovery Grant in the design of microarray experiments. He has also served extensively as a referee for many statistical journals and was the Chair of the Scientific Program Committee for the 15th Australian Statistical Conference held in July 2000. He is a member of the American Statistical Association, the Institute for Mathematical Statistics and the Statistical Society of Australia.

Timothy Staniford

TRC Mathematical Modelling, University of Adelaide

Tim Staniford graduated from the University of Adelaide in 2004 with a first class honours degree in applied mathematics and statistics. In 2005 he commenced working for TRC Mathematical Modelling at the University of Adelaide where he focussed on statistical analysis projects, for clients including DSTO. He worked on a diverse range of consultancy projects, including the modelling of highly complex two-dimensional probability distributions, clinical trial calculations, and an algorithm for the minimisation of waste in a foam-cutting procedure.

Michael Rumsewicz
TRC Mathematical Modelling, University of Adelaide

Michael Rumsewicz is Director of the Centre for Defence Communications and Information Networking at the University of Adelaide. His studies were undertaken at the University of Adelaide, graduating with a first class honours degree in Applied Mathematics in 1984 and awarded his PhD in 1989. Michael has worked primarily in the telecommunications industry and academia. From 1988-1994 Michael worked at Bellcore (USA), specialising in performance analysis of telecommunications systems. He subsequently returned to Australia and led performance analysis research at the Software Engineering Research Centre at Royal Melbourne Institute of Technology (1994-1999), where he developed new concepts on robustness and scalability for distributed web server platforms. From 1999 until 2001 he was with Ericsson Australia as the Team Leader for the Ericsson - Melbourne University Laboratory (EMULab) which specialised in network performance research. He returned to his alma mater in 2002, leading a range of network and statistical analysis projects for clients including Telstra, Tenix and DSTO. He has written over 40 refereed journal and conference articles, has been granted two patents, and has been a reviewer and programme committee member for a number of international conferences.

Oleg Mazonka
Davtec IT

Oleg is a senior software engineer at Davtec IT currently working as a contactor in Weapons Systems Division. He has a Ph.D in theoretical physics obtained from the Institute for Nuclear Studies in Warsaw in 2000. Oleg worked on software projects in different areas such as formal verification and validation tools, monitoring and control systems for embedded OSes, radar simulation, distributed systems for analysing geo-referenced video data. His professional interests include mathematics and C++.

Jeremy McMahon
TRC Mathematical Modelling, University of Adelaide

Jeremy McMahon graduated from the University of Adelaide in 2003 with a first class honours in Applied Mathematics. In 2008 he was awarded his PhD in Applied Mathematics, undertaking research in Markov Decision Processes. Both before and since undertaking his PhD, Jeremy worked at TRC Mathematical Modelling at the University of Adelaide as a Research Associate/Research Fellow, researching computer architectures supporting immersive audio environments for massively multi-player games, optimisation techniques for the manufacturing industry, undertaking consulting on statistical analysis projects, and developing state of the art telecommunications analysis planning systems.

Duncan Fletcher
Weapons Systems Division

Duncan is a senior weapons engineer with Weapons Systems Division of DSTO at Edinburgh in South Australia. He graduated from the University of Adelaide in 1998 with honours in Computer Systems Engineering. In his ten years at DSTO Duncan has gained extensive experience in the development of guided air weapon models, simulations and associated software. In 2002 he was posted to Dstl Farnborough in the United Kingdom Ministry of Defence to develop new capabilities in high fidelity guided weapon simulation. He is currently a lead engineer in WSDs Modelling and Simulation Software Development Cell (MOSSDEC) and Senior Engineer of the Range Safety Template Toolkit (RSTT) development projects. The MOSSDEC is responsible for developing modelling and simulation software in direct support of the ADF. The RSTT projects are currently developing Weapon Danger Area (WDA) generation capabilities for ASRAAM, JASSM and experimental hypersonic vehicles.

Michael Jokic
Weapons Systems Division

Michael is a research engineer with Weapons Systems Division of DSTO. He holds a Ph.D. in Aerospace Engineering from the University of Queensland where he also completed a Bachelor of Engineering degree (Mechanical and Space Engineering) with first class honours in 1999. Since starting with DSTO in 2005, Michael has been involved in the development of a range safety weapon danger area generation capability, and guided weapon simulation models. Michael's contribution to the Range Safety Template Toolkit (RSTT) project and DSTO Long Range Research program was recognised in 2008 with a DSTO Early Career Achievement Award.

Contents

Acronyms

EVT	Extreme value theory
GGM	Generic guided missile
IQR	Interquartile range
KDE	Kernel density estimate
MEB	Maximum energy boundary
MLS	Mean log scaled
PDF	Probability density function
RSTT	Range safety template toolkit
WDA	Weapon danger area

1. Introduction

1.1 Background and Purpose

In late 2004 the Defence Science and Technology Organisation (DSTO) of the Australian Department of Defence initiated development of an advanced, probabilistic range safety assessment system for guided air missiles. The resulting system is called the Range Safety Template Toolkit (RSTT). The key output of RSTT is a range safety template, which defines the evacuation area for a planned trial, also known as a weapon danger area (WDA) or weapon safety footprint area. WDA is the standard NATO term for such an area.

In designing RSTT, DSTO proposed a template generation methodology based on high fidelity Monte Carlo simulation of the missile, producing large sets of ground impacts for both nominal and off-nominal (i.e. failed) missile fly outs. One step in the proposed methodology required RSTT to generate two-dimensional ground impact probability density functions from the large sets of ground impact coordinates.

Early experimentation into probabilistic methodologies and Monte Carlo simulation showed non-Gaussian ground impact scatter was typical for guided weapon systems. The observed scatter, and limited time and computing resources became the primary constraints for the DSTO approach. The RSTT development plan therefore called for research and development (R&D) to be conducted into appropriate statistical techniques for the calculation of probability distributions, from minimal data sets. As a large project, not all R&D aspects of RSTT development could be handled internally by DSTO, so this particular task was contracted to a research centre of the University of Adelaide, the Centre for Defence Communications and Information Networking (CDCIN), formerly known as TRC Mathematical Modelling (TRC). The CDCIN team drew on statistics expertise from the University's Applied Mathematics department and regularly consulted with the DSTO RSTT development team on the direction of their research.

Three client reports were produced by CDCIN over a three year period from January 2005 to February 2008, proposing an initial solution to the problem and then examining various issues in the application of that solution. This DSTO Report presents for public release the consolidated analysis and findings of this R&D into the problem of calculating probability distributions from minimal data sets for range safety purposes.

1.2 Scope

Representative unclassified sets of simulated ground impact data were generated by DSTO and provided to the CDCIN team as the basis for their R&D activities. The activities undertaken by the team included the following:

Phase 1:

1. Analyse distributions to understand heterogeneous processes, developing an understanding of "typical" impact distribution data generated by DSTO.
2. Develop robust estimation methods: Investigated EVT (Extreme Value Theory) and other methods with representative distributions for evaluation of statistical method effectiveness.
3. Investigated database reduction and performance improvement methods: To make the RSTT as computationally efficient as possible, investigated database reduction and performance methods for later implementation in software.

Phase 2:

1. Investigated the convergence properties of probability density function estimation techniques with increasing size of dataset.
2. Investigated the degree of resolution required to provide impact distributions meaningful to use in RSTT.

Phase 3:

1. Identified issues associated with “clustering” of points and their impact on Kernel Density Estimation and bandwidth selection and provided a process for addressing these issues.

1.3 Report structure

This report is structured as follows:

- A table of acronyms precedes the Introduction
- Section 1, this section, introduces the report.
- Sections 2, 3 and 4 provide the outcome of the CDCIN activities undertaken in their three phases of work respectively, as described above in the Scope section above.
- Section 5 outlines subsequent analysis performed at DSTO into the application of non-diagonal bandwidth matrices.
- Section 6 provides a conclusion to the report and recommendations for further investigation.
- Appendix A lists the data provided to CDCIN for their first phase of work.
- Appendix B provides a series of plots to supplement Section 2.

The references appear following the conclusion.

2. Data Analysis and Development of Robust Estimation Procedures

In this section we report against the following, Phase 1, activities:

1. Analysed distribution to understand heterogeneous processes, developing an understanding of “typical” impact distribution data generated by DSTO.
2. Develop robust estimation methods: Investigated EVT (Extreme Value Theory) and other methods with representative distributions for evaluation of statistical method effectiveness.

The data generated by DSTO covered a range of missile launch scenarios and failure modes considered to be typical of the data that might be generated for actual weapons systems.

The completion of these activities involved the following steps:

1. Investigation of techniques suitable for generating approximate probability density functions modelling missile impact data.
2. For a given data set, investigation of techniques for generating "boundaries" on range sites that result in a probability of injury less than pre-determined safety levels.

2.1 Data Analysis

2.1.1 The Data

Weapons Systems Division of DSTO generated the data files listed in Appendix A, covering a range of different operational scenarios.

We define a *scenario* to be a set of simulation results derived using the same input data except for the type of failure, if any, that takes place and the seeds of noise sources. For example, two sets of simulation results corresponding to “no failures occurring” and “locking of fins at some time during the missile flight” are from the same scenario if the launcher altitude and velocity and target altitude and velocity are unchanged between the two data sets. If the launcher altitudes were different, the results would be said to derive from different scenarios.

For a given scenario and failure mode, the variables used from the data generated by DSTO contain information described in Table 1.

Table 1: Description of variables used in analysis

Variable	Description
SuccessfulInterceptTime	The expected time taken for a successful intercept of the target. Note that this column was added to each dataset by TRC based on the times given in the written documentation provided by DSTO. The times provided were to two significant figures.
Input: FailureTime	The (random) time at which the failure occurred during the flight of the missile.
Output: ImpactPointX	The x-coordinate of the impact point of the missile.
Output: ImpactPointY	The y-coordinate of the impact point of the missile.

In order to illustrate the proposed techniques of analysis of the missile impact data, these techniques will be demonstrated on the data provided for a particular scenario. The details of this scenario, as provided by DSTO, are given in Table 2. Unless otherwise stated, all plots provided in this section correspond to this scenario. Other scenarios were analysed.

Table 2: Details of the scenario to be used to illustrate the proposed techniques of analysis

Parameter	Value
Launcher and target altitude	1500 metres
Launcher and target speed	400 metres / second
Launcher flight direction	North
Target flight direction	South-West
Target location	8km North, 2km East of launcher location
Target manoeuvre	None

Figure 1 shows a scatter plot of 20,000 impact points for this scenario, where no failure occurs during the flight of the missile. These data points were provided in the file 'nofault_ggm_000_1-20000.csv'.

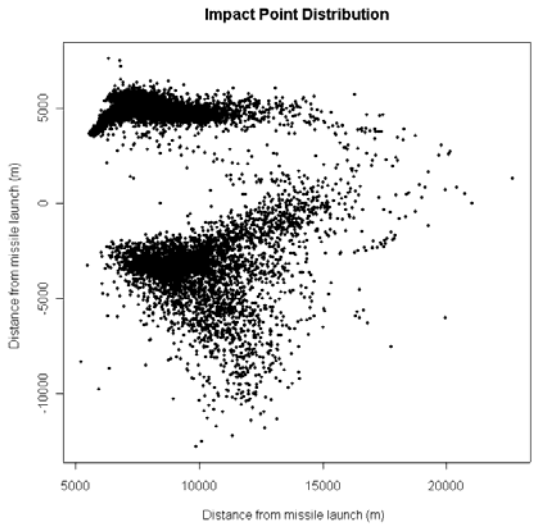

Figure 1: Impact distribution for scenario in Table 2 where no failure occurs (failure mode 0)

Figure 1 shows that there is a high degree of heterogeneity within the dataset, including two distinct areas of high impact density, and a considerable amount of dispersion of the impact points. Discussions with Duncan Fletcher and Robert Graham of Weapons Systems Division indicate that most of this dispersion is due to the missiles missing the target, and then turning around to re-attack the target. It can also be seen that the shapes of the areas of high density are somewhat irregular. A consequence of the high degree of heterogeneity and irregularity is that this data may not be well modelled by a mixture of bivariate normal distributions. It is of particular interest that for this scenario, the area of highest density occurs very near the boundary of the convex hull of the impact points. This may be important in the analysis of this data since it may cause the chosen method of density estimation to assign considerable probability density to regions where no impact points have been observed. This is discussed further in a later section of this report.

Figure 2 shows a scatter plot of 20,000 impact points for the same scenario, where at some point during the flight of the missile, all actuators lock to zero deflection. This failure occurs at a random time that is uniformly distributed on the interval 0-15 seconds. A certain subset of these impact points has been highlighted in blue. This is explained below. These data are provided in the file 'faultset_all_actuators_zeroed_ggm1-20000_6.csv'.

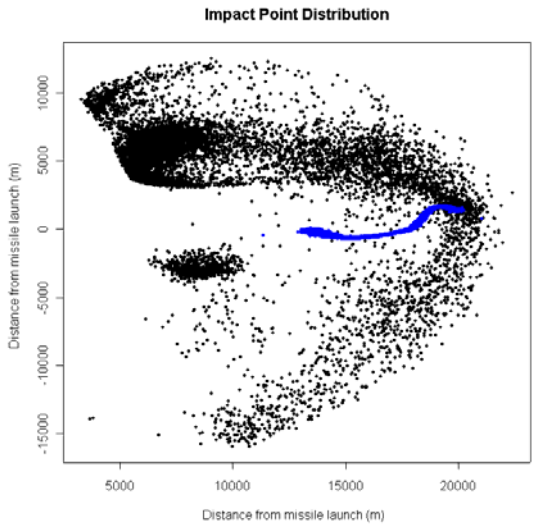

Figure 2: Impact distribution for scenario in Table 2 where all actuators lock to zero deflection (failure mode 1)

The main features of Figure 2 are similar to those of Figure 1, except that there is a narrow area of very high impact density around a segment of the launcher trajectory line. This area corresponds to the subset mentioned above that has been highlighted in blue. Such an area in a scatterplot of the impact points has been referred to as a "hook" in discussions with Weapons Systems Division. In this case, the hook appears to be contained entirely within the impact distribution, unlike the area of high density in Figure 1 which lies on the boundary of the distribution. There is also a higher degree of dispersion of the impact points in the zero deflection case than in the no failure case. This may be due to the additional variation introduced by the randomly generated failure time. As in Figure 1, the areas of high impact

density are irregular in shape, and this data may not be well modelled by a mixture of bivariate normal distributions.

It has been determined that the “hook” discussed above consists of cases in which failure occurs prior to the expected time of intercept. For this scenario, the expected time of intercept is 7.4 seconds. This hook has been extracted from the dataset according to the failure time, and is shown in Figure 3.

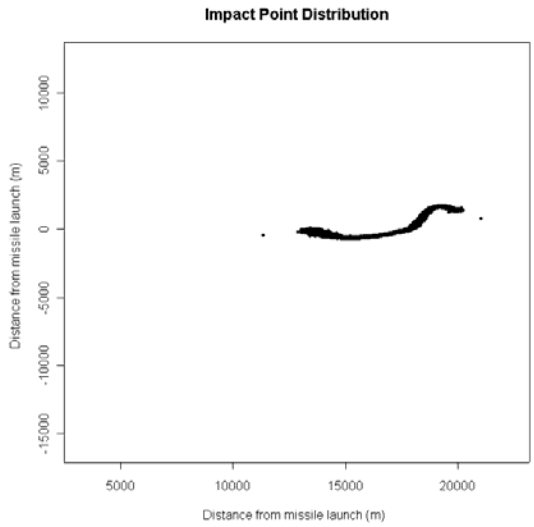

Figure 3: Impact distribution for pre-intercept failure times from Figure 2

Figure 4 shows the distribution of impact points when failure occurs after time 7.4 seconds.

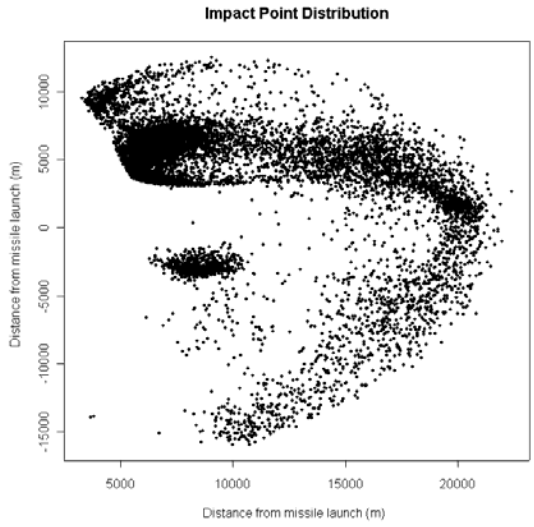

Figure 4: Impact distribution for post-intercept failure times from Figure 2

It is clear from Figure 3 and Figure 4 that the distribution of the impact points is very different when the failure occurs before the time of intended intercept than when it occurs after this time. Figure 3 shows that if the failure occurs before this time, the resulting distribution contains very limited dispersion. It appears that in this situation, the missile simply continues along a trajectory similar to its initial trajectory and eventually impacts the ground. The hook shown in Figure 3 is clearly a region of high impact density. If this hook is contained entirely within the envelope of the impact points shown in Figure 4, the chosen method of density estimation is likely to provide a reasonable estimate near the boundary of the impact distribution, as internal regions should not affect the boundary regions to any significant extent. However, as mentioned above, if the hook lies on or outside the boundary of this envelope, it may cause the chosen method of density estimation to assign considerable probability density to regions where no impact points have been observed. This will be discussed further at a later stage of this report.

2.1.2 Major issues in Data Modelling

From the descriptive analysis provided in the previous subsection, we see that the following major issues need to be addressed by any technique attempting to reasonably model data of the form provided:

1. The data are heterogeneous. In this case, there are clearly identifiable regions within the impact zone corresponding to different operational modes. For example, failure before intended intercept time can result in a tightly defined, highly correlated, impact zone. Failure after intended intercept time results in a much more widespread impact distribution, but still having distinct regions of higher density.
2. High density impact zones can occur at the edge of the “impact” envelope (see Figure 1), or more centrally in the impact envelope (see Figure 2).

For the purposes of developing a RSTT, high density zones central to the impact envelope may not have to be modelled with great precision. It is more important that the density estimate correctly estimates the overall probability of impact in those regions.

1. On the other hand, high density, sharply defined regions on the edge of the impact envelope will require a more careful treatment. Correctly recognising a sharp boundary may be useful in assessing the risk to immediately adjacent regions. However, the consequences of incorrectly identifying such a boundary could be severe and, for this reason, it is important to recognise the limitations of the available data. In particular, even if a sharp boundary is present in all simulated scenarios, it may not be possible to predict exactly how that boundary would be affected by small violations of the scenario assumptions as are bound to occur in practice. For this reason it would be prudent to allow for a margin well beyond that indicated directly by the data.
2. At this time we have not been provided with information on typical Maximum Energy Boundaries. A Maximum Energy Boundary defines the absolute maximum range of a missile, in any given direction, as limited by issues such as launch altitude, velocity, weight and fuel load. Thus, any models developed for the impact distribution of a missile should be flexible enough to adhere to external limits imposed by a specified Maximum Energy Boundary.

3. Finally, in considering models for the data and their interpretations, it is essential to note that any such modelling assumes the data to be representative of the behaviour of the physical system. In particular, the variation apparent in the data is due solely to variation in certain input parameters such as failure time, wind-speed and seeker noise. The extent to which the data is representative is determined by the extent to which the variation of the inputs represents the system being modelled. Throughout this report it is assumed that the data provided are representative of the system intended by DSTO. Although this assumption is critical, by its nature it cannot be tested on the basis of the data alone. Therefore it is an assumption rather than a conclusion of this report that the data are representative of the system intended by DSTO.

2.2 Techniques for generating PDFs

In the preliminary phase to this project, TRC Mathematical Modelling provided a report on the use of Extreme Value Theory and its potential application to determining regions encapsulating a given percentage of likely impact points [1]. Under this approach, precise modelling of the entire impact area is not required as the focus is on the "edge" of the impact zone and how far it might extend in any particular direction. In this phase of the project, the emphasis has changed to enabling the RSTT to compute the expected risk of harm for a given test template and map of population density. This requires modelling the probability density function of impacts across the entire impact area and therefore precludes the use of EVT.

In the context of the missile impact distribution, the associated probability density function assigns to any impact point (x,y) a probability density. An intuitive way of thinking of the probability density function is as the continuous analogue of the probability mass function. A higher probability density in a particular region indicates a higher probability of the missile landing in that region. The integral of a probability density function over all possible impact points (x,y) is 1. It should be noted, however, that in order to perform the calculations necessary for this report, such as the calculation of the expected number of casualties, the estimate of the probability density function is calculated over a discrete grid of values. Therefore, the density estimate used in the calculations throughout this report is in fact a probability *mass* function rather than a probability *density* function. It should also be noted that during the process of generating this probability mass function, it is normalised so that the sum of the probabilities over all grid squares is equal to 1. Thus, all density estimates used in the analysis throughout this report are valid probability mass functions. This information, when combined with information on other external factors, such as population density, can then be used to generate estimates of overall injury rates.

2.2.1 Application of Kernel Smoothing to PDF Generation

The problem at hand is to estimate the probability density function for the impact distribution across the area of interest. There are many approaches to density estimation and TRC Mathematical Modelling has focussed on Kernel Smoothing (see [2]) for this application. In broad terms, methods for density estimation can be classified either as parametric or non-parametric.

Parametric methods rely upon *correctly* specifying a family of distributions, such as the bivariate Gaussian, and then adjusting the parameters of that distribution to fit the data. Although parametric methods could be expected to provide the highest statistical efficiency, they are not applicable in this case because of the complexity of the data. In particular, none of the available families of parametric densities is suitable.

Non-parametric methods include histogram methods, kernel density estimation and various series expansions. Unlike parametric estimation, non-parametric methods do not assume a particular form for the impact distribution. Histogram methods are extremely simple to implement and involve the fewest assumptions about the impact distribution. However, they do not smooth or interpolate between points and hence are not suitable. Kernel density estimation and series expansions both provide for smoothing and interpolation between data points. It was decided to focus on kernel density estimation for practical reasons; namely, the method has been extensively studied [[3, 4] and other references], its theoretical properties are well understood and efficient implementations are widely available.

Kernel Smoothing works in the following way (taken from [2]):

A probability density function f of a random variable X can be defined by

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h),$$

where h is a constant. For any given h, we can estimate $P(x - h < X < x + h)$ by the proportion of the sample falling in the interval $(x - h, x + h)$. Thus, a natural estimator of the density is given by

$$\hat{f}(x) = \frac{1}{2hn} \times \text{number of observations falling in } (x - h, x + h).$$

This is known as the *naïve estimator*, and can be expressed more generally as

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \times w\left(\frac{x - X_i}{h} \right),$$

where w(·) is a weight function given by

$$w(x) = \begin{cases} \frac{1}{2} & \text{if } |x| < 1, \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to generalise this estimator by replacing the weight function w(·) with a *kernel function* K(·) that satisfies the condition

$$\int_{-\infty}^{\infty} K(x) dx = 1.$$

Usually, K(·) will be a symmetric probability density function. Thus, the *kernel estimator* with kernel K(·) is defined by

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where h is known as the *bandwidth parameter*. The idea of the kernel density estimator is that it smooths the raw data into a probability density function. That is, it spreads the weight of each data point over a wider area (determined by the kernel function) so that it "fills in" the gaps between each data point, whilst assigning greatest probability density to the areas with the greatest concentration of data points.

The naïve estimator can be considered as the sum of equal-sized 'boxes', with each box centred at an observed data point. In the same way, the kernel estimator can be considered as a sum of smooth 'bumps' placed at the observations. The kernel determines the shape of the bumps while the bandwidth h determines their width. If the selected bandwidth is too narrow, these bumps will not overlap, and the resulting density estimate will be a collection of isolated bumps of probability density. This may also occur if the data are too sparse. On the other hand, if the bandwidth is too large, each point will be spread over a very large area, and considerable amounts of probability density will be allocated to areas where no data has been observed. Thus, it is important to select an appropriate bandwidth.

The 2-dimensional generalisation of the kernel estimator is given by

$$\hat{f}(x, y) = \frac{1}{nh_x h_y}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_x}, \frac{y - Y_i}{h_y}\right),$$

where K(·) is now a 2-dimensional kernel function. Clearly, there are a number of things to be determined before using the kernel density estimator. Firstly, one must choose the kernel function K(·). Secondly, in practice, the density estimate is computed over a discrete grid by evaluating the formula given above at each point on the grid. Thus, it is also necessary to determine the dimensions of this grid. Throughout this report, this will be referred to as the granularity of the density estimate. Finally, and importantly, one must choose the values of the two bandwidth parameters, h_x and h_y. In selecting the bandwidths, there are a number of methods from which to choose. For 2-dimensional kernel density estimation, the most common methods of selecting the bandwidth are to apply a certain standard-distribution based formula or to use cross-validation. These issues are addressed in Section 2.2.3.

2.2.2 Application of Kernel Smoothing to the Missile Impact Data

By applying kernel density estimation to the missile impact data and performing certain manipulations on the kernel density estimates obtained, it is possible to obtain an estimate of the missile impact distribution that appears to be reasonable.

This process has been carried out for the scenario described in Table 2, and a final estimate of the missile impact distribution was obtained. A representation of this final estimate is shown in Figure 5, below. The figure shows the probability mass function obtained through:

- Kernel smoothing the data provided.
- Applying an artificially created (circular) Maximum Energy Boundary to the probability mass function created.
- Using different colours in the figure to represent different levels of probability density.

The legend at the left of the plot gives an indication of the level of density to which the various colours correspond. Note that due to the large number of grid squares into which the area of the density is divided, the levels of probability assigned to each grid square are very small. For this reason, the plots of the kernel density estimates in this report show the log (base 10) of the density. Thus, for example, a grid square of the same colour as the first square on the legend indicates an estimated probability density between 0.01 and 0.001. Areas with estimated density less than 10^{-20} have simply been coloured white. The apparent truncation of the probability density estimate in the top left and bottom left regions is due to application of the circular Maximum Energy Boundary used in this example. It is important to note that the kernel density estimate plots shown in this report depict the probability density per unit-cell. To calculate the probability per metre-squared the probability density in each cell must be divided by the area of the cell.

The step-by-step process by which this estimate was obtained is now explained.

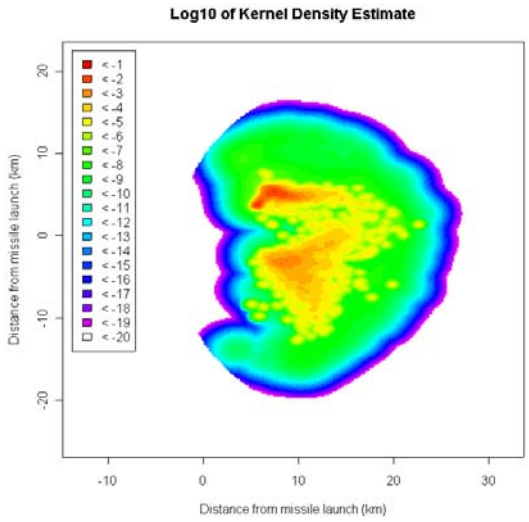

Figure 5: Final kernel density estimate for the scenario described in Table 2

The first step toward obtaining an overall kernel density estimate for a particular scenario is to compute individual kernel density estimates for each failure mode for that scenario. Figure 6 and Figure 7 show the kernel density estimates for the data shown in Figure 1 and Figure 2, respectively, which correspond to failure modes 0 (no failure) and 1 (zero deflection) for the

scenario described in Table 2. The actual impact points are also shown in order to illustrate the quality of the density estimate.

For the kernel density estimates shown in Figure 6 and Figure 7, the bandwidth parameter was chosen according to a formula discussed in detail in Section 2.2.3.3 and the kernel estimator was applied on a 256×256 grid overlaid on the dataset. The area for which the density estimate was computed for this scenario was determined from the range of the combined data for both failure modes. The individual kernel density estimates for each failure mode of that scenario were computed on a common grid in order to be able to calculate a weighted average of the distributions of the various failure modes for that scenario. This weighted average is discussed and calculated in Section 2.3.1, below. The overall x-range of the density estimate for the scenario to which Figure 6 and Figure 7 correspond is (-6727.9, 32693.8), the overall y-range is (-25956.4, 22538.2), and the dimensions of the rectangles in the discretisation of the impact distribution for the individual failure modes are 153.99 m by 189.43 m. This will also be the discretisation of the weighted average.

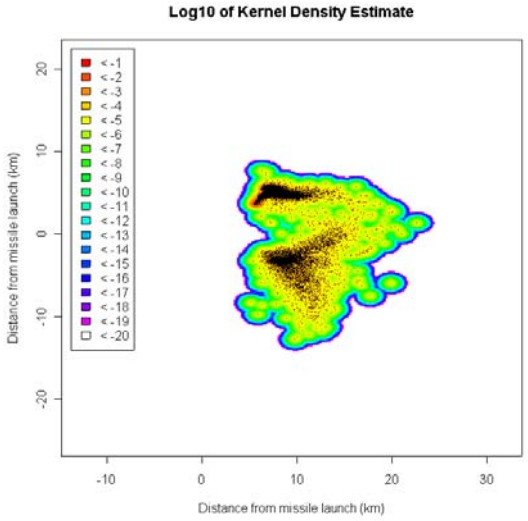

Figure 6: Kernel density estimate for scenario in Table 2 where no failure occurs (failure mode 0)

It can be observed from Figure 6 that the kernel density estimate appears to be a reasonable estimate of the underlying density that generated the impact points shown. Since the vast majority of the impact points fall in one of two areas of especially high impact density, the estimated probability density in these two areas is much greater than in the remaining area of the plot. A further result of the two areas of very high impact density is that the bandwidth chosen is quite small, and therefore the tails of the kernel density estimate decay quite rapidly.

As discussed above, a possible consequence of not having enough data or choosing too small a bandwidth is that the resulting kernel density estimate is a collection of isolated bumps of probability density, corresponding to the data points. In Figure 6, above, the kernel density estimate appears to be a reasonable estimate in most areas of the plot, but in the areas where the data are sparse, isolated bumps of density can be observed. The issue of bandwidth

selection is further discussed below, as is the issue of assessing whether or not the dataset is of sufficient size.

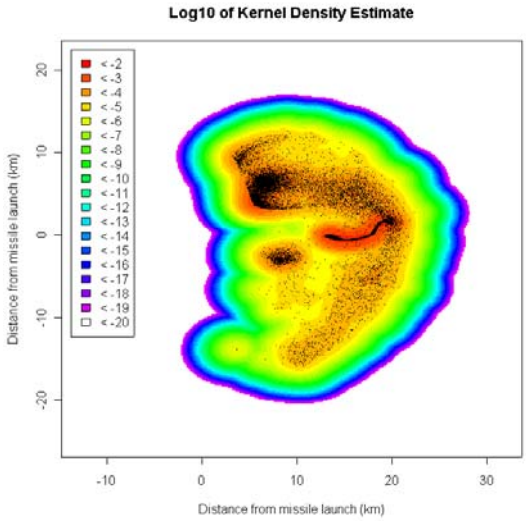

Figure 7: Kernel density estimate for scenario in Table 2 where all actuators lock to zero deflection (failure mode 1)

The kernel density estimate in Figure 7 also appears to be a reasonable estimate of the underlying density that generated the impact points shown. The areas of greatest estimated probability density correspond closely to the areas of greatest impact density. The dispersion of the impact points is greater in the zero deflection case than in the no failure case. Consequently, the bandwidth chosen is larger, and the tails of the kernel density estimate decay less rapidly in the zero deflection case. However, the rate of decay in the zero deflection case is still quite rapid since this is an inherent feature of the bivariate normal kernel.

The code used to analyse the data provide by DSTO and to generate the plots in this report was developed by TRC Mathematical Modelling in the "R" statistical analysis package. R is a freeware software package available for download from www.r-project.org. It has built-in routines for handling large datasets and applying kernel smoothing techniques to 2-dimensional data, such as the impact point data generated by DSTOs missile flight simulations.

Looking at the figures, we observe the following key features:

1. As described earlier, kernel density estimation "fills in" areas of low impact density that exist physically between two high density impact regions. See Figure 7 for an example. The amount of in-fill, or spreading, is controlled by the bandwidth parameter. The following section provides a detailed coverage of how the bandwidth parameter affects the probability density function generated.
2. When a high density impact area exists on the edge of the envelope of the impact points, kernel smoothing spreads a portion of the distribution to the area outside of the envelope. See Figure 6 for an example. If the simulated data are believed to provide a very accurate representation of the actual impact distribution that could be

expected in practice, there would be an argument for applying a smaller bandwidth parameter to points in those impact regions. In effect, this would "tighten" the probability density function generated to more closely represent the data. On the other hand, if there is any doubt about the integrity of the input data, then such a tightening would be a more aggressive, rather than conservative, approach to the development of range safety templates.

3. Another issue to consider in determining the effectiveness of kernel density estimation is the size of the dataset. Clearly, if the number of data points available is too small, it is impossible to obtain a good estimate of the underlying density. As mentioned earlier, a possible consequence of this is that the resulting kernel density estimate consists of a collection of isolated bumps of probability density each corresponding to a data point. This issue is further complicated by the fact that the given data exhibit a high degree of heterogeneity in the density in different areas of the impact distribution. Thus, while there may be plenty of data available to estimate the density in some areas of the impact distribution, the data may be very sparse in other areas which do in fact contain a significant amount of probability density.

Items 1 and 2, in particular, highlight that while kernel smoothing has many useful characteristics with regards to generating probability density functions from simulated data, its use in the development of range safety templates requires careful application and the review of input data by appropriate subject matter experts in the weapons area. The selection of appropriate kernel smoothing input parameters is not simply a matter of statistical analysis and blind application of existing formulae for selection of the bandwidth parameter.

2.2.3 Issues in the Application of Kernel Smoothing

2.2.3.1 Choice of Kernel

As mentioned in Section 2.2.1, there are a number of decisions that need to be made in order to apply kernel density estimation. The first of these is the choice of the kernel function itself. A common choice of bivariate kernel function is the bivariate normal density function. There are a number of other kernel functions that may be used, some of which are more computationally efficient than the normal kernel. However, computational efficiency has not presented any significant problems in the implementation of the kernel density estimator.

A second issue with the normal density function is that the tail decays faster than e^x. Thus, the tails of the resulting kernel density estimate also decay exponentially. In the case of the missile impact distribution, the interpretation of this is that as the impact location moves away from the regions of high impact density, the probability density associated with that impact location decays very rapidly. This may be an undesirable feature of an estimate of the impact density, since it may imply that impact locations just outside the envelope of the impact points observed in the given dataset have very small associated impact density. It may be more desirable to associate higher impact densities with these points. However, this issue is also related to the choice of bandwidth, since a larger bandwidth creates a longer tail. The issue of bandwidth selection is further discussed below, but for now, it is sufficient to note that rapid decay of the tail of the normal kernel can be overcome, to some extent, by an appropriate choice of bandwidth.

One advantage of the normal kernel in applications such as the range safety project is that it is much more likely to be familiar to non-statisticians, and its properties are much more widely known. Since the two major issues with the normal kernel can both be overcome, the kernel function used for the analysis throughout this report is the bivariate normal density function.

2.2.3.2 Granularity and Range

A second aspect to be determined is the grid over which the density estimate is computed. In order to determine this grid, it is sufficient to determine the range over which the density estimate is to be computed and the dimensions of the grid (that is, the granularity).

For the analysis throughout this report, the range over which the density estimate is computed has been determined by taking the range of the impact point data and adding a border of width 10km around the outside. In all of the kernel density estimates shown throughout this report, it can be seen that the probability density at any point outside this range is less than 10^{-20}. Therefore, it has been decided that this is a sufficiently wide range over which to compute the kernel density estimate. In a subsequent step of the procedure for obtaining an estimate of the impact density for a particular scenario, a maximum energy boundary will be applied to the kernel density estimate. The application of a maximum energy boundary sets the estimated probability density in all grid squares outside of the boundary to 0. Therefore, provided that the range of the original kernel density estimate is large enough to contain the entire maximum energy boundary, the effect of the range on the kernel density estimate will be negligible. In fact, x- and y-ranges of the maximum energy boundary may be appropriate choices for the range of the grid over which to compute the kernel density estimate. Maximum energy boundaries are further discussed in Section 2.3.2.

In determining the granularity of the estimate, there are two competing factors to consider. Firstly, it is clear that a finer discretisation will provide a finer estimate of the density function. However there are also limits imposed on the granularity by computational issues. For the analysis described in this report, the granularity was chosen such that a sufficiently fine estimate was obtained whilst maintaining a reasonable computation time. The granularity of each density estimate given in this report is 256×256. However, the software used does allow the user to select a granularity of their own choice (for example, 128 or 512).

2.2.3.3 Bandwidths

The final choice to be made is the values of the bandwidth parameters. In many ways, this is the most important choice since “both theory and practice suggest that choice of kernel is not crucial to the statistical performance of the method and therefore it is quite reasonable to choose a kernel for computational efficiency” [3]. The most common methods of selecting the bandwidth are to use one of two common standard-distribution based formulae, or to use cross-validation. These standard-distribution based formulae are referred to in the literature as “rules of thumb”. Although this choice of term suggests that the formulae may not be appropriate for range safety purposes, they are in fact quite valid as long as certain assumptions about the data hold. See section 3.4.2 of reference [3] for details.

The rules of thumb given in the literature actually apply to 1-dimensional datasets. They can be applied separately to each dimension of a 2-dimensional dataset, but this is not always

appropriate. The first rule of thumb is to select the bandwidth according to the following formula:

$$h = 1.06 \times \min\left\{\sigma, \frac{\text{IQR}}{1.34}\right\} \times n^{-\frac{1}{5}},$$

where IQR is the interquartile range of the distribution.

A theoretical derivation of this rule is given in [2, pg 45]. The second rule of thumb is a simple variation of the first, where the factor of 1.06 is replaced by 0.9. The motivation behind this variation is that with a factor of 0.9, the error of the density estimate will be within 10% of the minimum error. However, it is more common to use the first rule of thumb.

The objective of the cross-validation methods is to choose the bandwidth that gives the best fit to the data. The data are partitioned into a number of equal-sized subsets. One of the subsets is removed from the dataset, and the remaining data are used to calculate a kernel density estimate. This is repeated for each subset, and a measure of the error of the estimate based on the removed subset is calculated. This procedure is repeated for a number of different bandwidths, and the bandwidth is chosen to give the minimum error. This is clearly a useful method of choosing the bandwidth. It also has the advantage that it may be used to select the best 2-dimensional bandwidth, rather than choosing the bandwidth independently for each dimension. However, it is considerably more difficult to implement, and if reasonable results can be obtained using a simpler method, it may not be necessary to expend the additional effort.

For the given data, the first rule of thumb appears to give reasonable results. Since it is the most common method, and the simplest to implement, the bandwidths used to obtain the kernel density estimates throughout this report have been chosen using the first rule of thumb.

A further possibility in bandwidth selection is known as an *adaptive bandwidth*, or *dynamic bandwidth*. The idea is that different bandwidths may be used for different regions of the (x,y) area of the density estimate. A number of dynamic bandwidth techniques have been developed, but to implement such a method effectively would take a considerable amount of additional implementation and validation. To effectively implement such a method is therefore not feasible within the timeframe of the current project.

2.2.3.4 Size of Dataset

Before applying kernel density estimation, it is necessary to ensure that the dataset contains a sufficient number of points to obtain a reasonable density estimate. The smallest dataset provided by DSTO from which a kernel density estimate was to be obtained was of 10,000 points. However, unless stated otherwise, the kernel density estimates shown throughout this report are based on datasets of 20,000 points. A possible way of determining the number of points required to obtain an acceptable kernel density estimate is to obtain independent samples of various sizes, and generate kernel density estimates for each. Assuming that the largest sample is sufficient to obtain an accurate kernel density estimate, the kernel density estimates based on the smaller samples can be compared to that of largest sample. If the

kernel density estimates based on samples of a certain size are sufficiently similar to the kernel density estimate based on the largest sample, and little is gained by using a sample larger than this size, then this may indicate that this is a sufficient number of data points to obtain an acceptable kernel density estimate.

An approach similar to that described above has been applied to a dataset of 50,000 points. However, with only 50,000 points, it is impossible to obtain many independent samples. For this reason, the samples used in this investigation are subsets of the 50,000-point dataset. Thus, the corresponding kernel density estimates may be more similar to the 50,000-point estimate than would be expected if the samples were independent. However, it is still possible to gain some insight into the number of points required to obtain an accurate kernel density estimate.

For the scenario described in Table 2, only 20,000 points were available. However, for a different scenario, a dataset of 50,000 points was available. Therefore, the dataset of 50,000 points has been used in this investigation. In fact, the scenario from which the dataset of 50,000 points was derived is a reflection in the x-axis of the scenario described in Table 2. Consequently, a degree of symmetry can be observed between both the scatterplots and kernel density estimates for these two scenarios. This symmetry will be discussed later in this report.

Firstly, the kernel density has been computed for the dataset of 50,000 observations. For samples sizes 5000 – 30,000 points, at intervals of 5000, 20 subsets of each size have been randomly selected, and a kernel density estimate has also been generated for each subset. The scatterplot and kernel density estimate for the entire dataset of 50,000 observations, along with typical scatterplots and kernel density estimates for each sample size are given in Appendix B. Appendix B also contains plots of 10^{-6} exclusion zones for each of the kernel density estimates.

For each size, the 20 subsets of that size have been used to compute an ‘average’ kernel density estimate for that size. This gives an indication of the ‘average’ kernel density estimate that might be calculated from a sample of that size (assuming that the sample of 50,000 is a good representation of the impact distribution). Each average kernel density estimate was then compared with the 50,000-point kernel density estimate.

The measure of difference used for the comparison of the average kernel density estimates with the 50,000-point estimate was the sum of the absolute differences at each grid square. It is not difficult to see that for two identical distributions, this measure will be equal to 0, and as two distributions become more and more different, this measure will increase. Throughout this report, the difference between an average kernel density estimate and the 50,000-point estimate using this measure will be referred to as a *density difference*.

Table 3 shows the density differences for average kernel density estimates for various sample sizes. Table 3 also shows both the absolute and relative marginal decrease in density difference as the sample size increases. These quantities give an indication of the improvement in the kernel density estimate gained by using a larger sample. For convenience, the absolute and relative differences have also been plotted against the number of points, and are shown in Figure 8 and Figure 9.

Note that whilst both the ratio of density differences and the relative decrease in density difference are both relative to the density difference with a further 5000 points, they do not represent the same quantity. The formulae by which each of these quantities was calculated are given below.

$$Ratio\ of\ \ Density\ Difference = \frac{Density\ Difference(n)}{Density\ Difference(n-5000)}, \qquad and$$

$$Relative\ Decrease\ in\ Density\ Difference = \frac{Density\ Difference(n-5000) - Density\ Difference(n)}{Density\ Difference(n-5000)},$$

where *Density Difference(n)* represents the density difference for an *n*-point average kernel density estimate. Whereas the ratio of density differences gives the size of the difference relative to the difference with 5000 fewer points, the relative decrease in density difference gives the decrease in density difference achieved by adding a further 5000 points. In fact, it can be seen from the formulae above that these two quantities are related by the equation

$$Ratio\ of\ \ Density\ Difference = 1 - Relative\ Decrease\ in\ Density\ Difference.$$

Table 3: Differences between average kernel density estimates and 50,000-point kernel density estimate for various sample sizes

Sample Size	Density Difference from 50,000-point estimate	Absolute Decrease in Density Difference	Ratio of Density Difference	Relative Decrease in Density Difference
5000	0.30563153	NA	NA	NA
10,000	0.2134398	0.09219173	0.698356613	0.301643387
15,000	0.1582162	0.0552236	0.741268498	0.258731502
20,000	0.11973258	0.03848362	0.756765616	0.243234384
25,000	0.09010242	0.02963016	0.752530514	0.247469486
30,000	0.06645264	0.02364978	0.737523365	0.262476635

In considering these results, it should be noted that the difference between the average kernel density estimate and the 50,000-point kernel density estimate do not necessarily reflect the size of the difference that could be expected for a kernel density estimate obtained from a single sample of the given size. In particular, it could be expected that the errors associated with a single sample would be larger that those associated with the average of 20 independent samples. It is, nevertheless, reasonable to assume that the general pattern of the difference decreasing as sample size increases would also apply to single samples. The magnitude of the difference for a given sample size could, in principle, be estimated from the 20 samples but this calculation has not been performed with the present data.

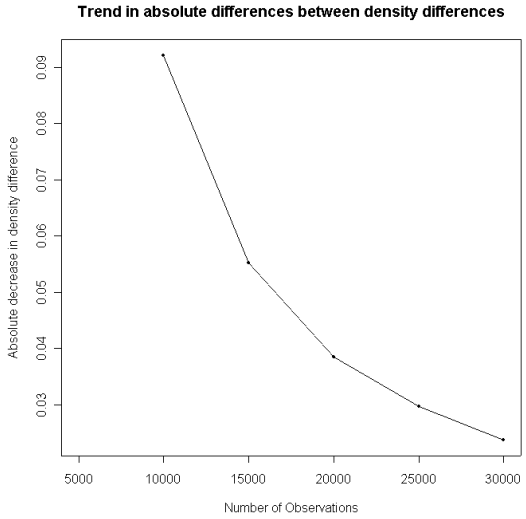

Figure 8: Trend in absolute marginal density difference as sample size is increased.

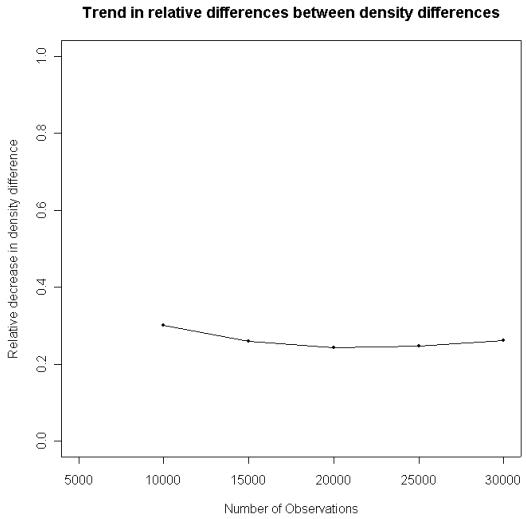

Figure 9: Trend in relative marginal density difference as sample size is increased

Together with Table 3, Figure 8 and Figure 9 show that as the sample size increases, the difference between the 50,000-point kernel density estimate and the average kernel density estimate becomes smaller. This is as expected. It can also be observed that as the sample size increases, the absolute marginal difference gained by adding extra points diminishes, and the relative marginal difference remains reasonably constant. It is difficult to determine an acceptable value of this difference, especially since the samples used were subsets of the 50,000 point dataset, but it would appear that the kernel density estimate may be significantly improved by increasing the number of points, even beyond 20,000.

This issue requires further investigation before undertaking a large project using kernel density estimates and the various techniques proposed in this report. Additionally, this issue

is further complicated by the high degree of heterogeneity in the density in different areas of the impact distribution. This issue is therefore also related to the choice of bandwidth since it may be appropriate to use a larger bandwidth for areas where the data are more sparse. Thus, in future analysis of the missile impact data, it may be helpful to use a dynamic bandwidth method.

It is important to note, however, that while there may be a relatively significant change in the details of the probability density function as the number of points is increased, this is but one measure of comparison and may not be the most significant measure.

Appendix B, Figure 64 through Figure 69, shows the convex exclusion zones generated for each of the datasets. Visually, we see that the exclusion zone changes very little as the number of points is increased beyond 15,000, and even the difference between 5,000 data points and 30,000 data points is relatively small. Based on this form of measure and this particular dataset, even 5000 points may be sufficient depending on the degree of accuracy required, especially if other safety margins will be subsequently applied.

2.2.3.5 High density Regions on Boundary of Impact Envelope

A further issue associated with the heterogeneity of the impact density is that if there is an area of high impact density that lies on the boundary of the impact envelope, it is possible that a considerable amount of the density associated with that area could be spread to regions beyond the impact envelope, where no data has been observed. However, this may not be as great a problem as it may seem, because Figure 6 and Figure 7 show that in the no failure case (Figure 6), where the areas of high impact density are much nearer to the boundary of the impact envelope, the amount of probability density spread to regions beyond the impact envelope is actually smaller. The explanation of this is that in the no failure case, smaller bandwidths are chosen, and the probability is spread over a smaller area. In combination with the relatively low probability of a failure as compared with the “no failure” mode, the actual amount of probability assigned outside of the simulated impact envelope in the merged probability density function is commensurately decreased. Figure 6 and Figure 7 show that the impact points are much less dispersed in the no failure case than in the zero deflection case.

The bandwidths used to generate the kernel density estimates shown in Figure 6 and Figure 7 are given in Table 4.

Table 4: Bandwidths used to generate kernel density estimates in no failure and zero deflection cases

Case	Bandwidth (x)	Bandwidth (y)
No Failure Case (Figure 6)	881.6	668.7
Zero Deflection Case (Figure 7)	2922.6	2294.3

The evidence from Figure 6 and Figure 7 suggests that the location of the high density regions in failure mode cases relative to the impact envelope is not the primary factor in determining the amount of probability density assigned to regions outside the impact envelope. The evidence suggests that the values of the bandwidths have a much greater influence on this.

From the discussion in this section, it is clear that there are many factors to be considered in generating an accurate kernel density estimate. There can be no "blind" process for producing a kernel density estimate that is appropriate for any dataset that may arise. It is necessary to ensure that appropriate statistical inputs have been used, and that the results obtained are consistent both with the data and with what might reasonably be expected in reality. Therefore, any procedure for generating a kernel density estimate should be reviewed by a panel of experts, including both statisticians and weapons experts.

2.3 Creating overall PDFs for a given scenario

2.3.1 Generating overall PDFs from individual failure mode PDFs

In the previous sections we described the process of generating a numerical probability density function from a given data set. For the purposes of the RSTT it is necessary to generate probability density functions for a given scenario. Recall that a scenario is made up of information covering all known failure modes, together with the "no failure" case, for a given set of other input parameters.

For the illustration of the technique of generating a density estimate for a particular scenario by an appropriate combination of the density estimates for the individual failure modes, TRC Mathematical Modelling has incorporated only two failure modes. That is, no failure (failure mode 0) and zero deflection (failure mode 1). The reason for this is that there was only one scenario for which data corresponding to multiple failure modes were available (namely, the scenario used throughout this report). Therefore, there was no other scenario that could have been used to illustrate the techniques proposed in this report. The data available for this scenario covered only failure modes 0 and 1. There was a dataset provided that corresponded to a further failure mode (single actuator freeze - failure mode 2), but this dataset did not correspond to the same scenario as the data for failure modes 0 and 1. In theory, it is straightforward to incorporate any number of failure modes with this technique, provided that an expert panel of some form is able to provide information on the probability of each failure mode.

Denote by p_i the probability of failure mode i, $i = 1, \ldots, N$, and let p_0 be the probability that no failure occurs. Then,

$$p_0 = 1 - \sum_{i>0} p_i .$$

Let PDF_i^S denote the probability density function for failure mode i, $i = 0, \ldots N$, of a given scenario S, where N is the number of failure modes. PDF_i^S is a matrix whose $(x,y)^{th}$ element is the probability of missile impact in grid square (x,y) given failure mode i of scenario S occurs. The probability density function for a given scenario, PDF^S, is given by

$$PDF^S = \sum_i p_i PDF_i^S .$$

As an example, for the data examined in Section 2.2.2, suppose we have:

Failure Mode 1: Fins locking to zero deflection, probability $p_1 = 0.0001$.

Probability of no failure, $p_0 = 1 - p_1 = 1 - 0.0001 = 0.9999$.

Combining the kernel smoothing generated probability density functions for these two modes, according to their respective probabilities of occurrence, we obtain the probability density function shown in Figure 10, below.

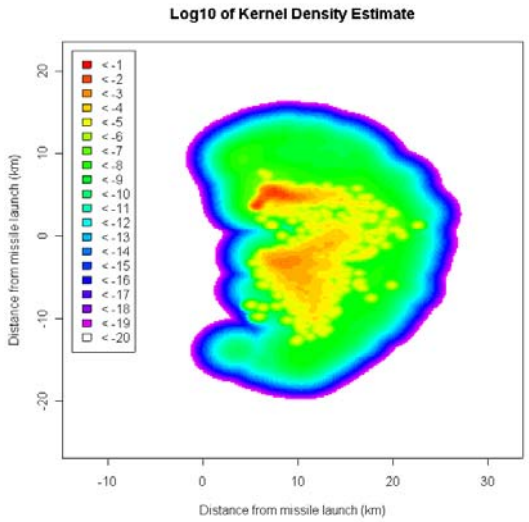

Figure 10: Overall PDF for scenario in Table 2

The interior of the combined PDF is very similar to that of the no failure PDF (Figure 6). This is clearly a sensible result since this case contributes the vast majority of the probability weighting. However, it can be seen from Figure 10 that the tails of the combined PDF are very similar to the zero deflection PDF (Figure 7). This is because the tail decays much more rapidly in the no failure case, as discussed earlier. Thus, in the tail regions the zero deflection PDF has much greater density, and hence contributes more to the weighted average, even though it is weighted by a probability of only 10^{-4}.

2.3.2 Generating overall PDFs from individual failure mode PDFs and using the Maximum Energy Boundary

The probability density functions generated in the previous sections were created by mixing 2-dimensional Gaussian distributions (since the kernel function chosen was the bivariate normal density). Recall that the general idea is to take each individual impact point and "spread" it over a wider area, with greatest concentration at the impact point itself. However, the normal density has non-zero probability at any point (x,y). Therefore, this results in a non-zero probability density even at grid squares huge distances from the impact point itself. This is unrealistic in practice.

Weapons Systems Division of DSTO has models for generating a "Maximum Energy Boundary" (MEB) for given test scenarios. The Maximum Energy Boundary is the maximum

distance, in any given direction, that a missile might travel given various factors, including, for example, fuel load of the missile.

Clearly, no impact point should lie outside the maximum energy boundary. However, the models used to generate both the impact points and the maximum energy boundary may not be perfect. Therefore, in the event that one or more impact points lie beyond the maximum energy boundary, it would be desirable for the software to be able to detect and report this. This has not been implemented in the current analysis.

If we are provided with an MEB for a given missile and test scenario, we can combine this information with the overall PDF generated in the previous section to form a potentially more realistic probability density function.

Denote by $MEBPDF^S$ the MEB constrained PDF for scenario S.

Denote by MEB^S the indicator matrix defining the MEB over the same grid layout used to define PDF^S. The (x,y) co-ordinate of MEB^S is 1 if grid square (x,y) is within the MEB and 0 otherwise.

Then the total probability mass of PDF^S that falls within the MEB is given by

$$p_{MEB} = \Sigma_x \, \Sigma_y \, PDF^S(x,y) \, . \, MEB^S(x,y),$$

and

$$MEBPDF^S(x,y) = \begin{cases} \dfrac{PDF^S(x,y)}{p_{MEB}} & \text{if } MEB^S(x,y) = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 11 shows an artificial MEB and Figure 12 shows the corresponding MEBPDF based on PDF shown in Figure 10. The software used to generate the MEBPDF shown in Figure 12 allows the user to specify an appropriate maximum energy boundary. The form of the MEB input in the function is of a grid of 0s and 1s, where a 1 indicates that a grid square is within the maximum energy boundary, and a 0 indicates otherwise. Therefore, in order to apply a realistic maximum energy boundary, it is necessary to generate such a maximum energy boundary and convert it to the appropriate form for the input to the function.

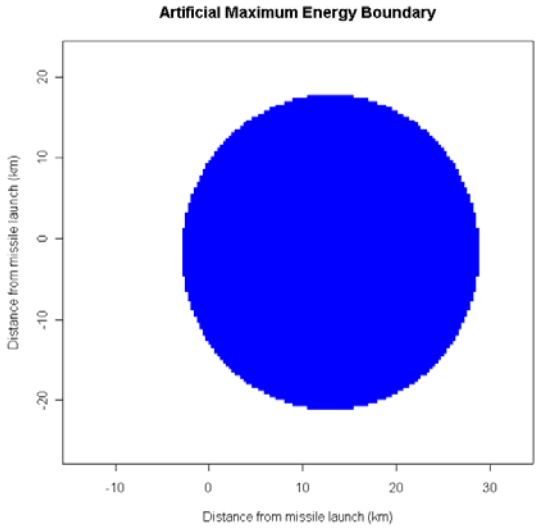

Figure 11: Artificial maximum energy boundary.

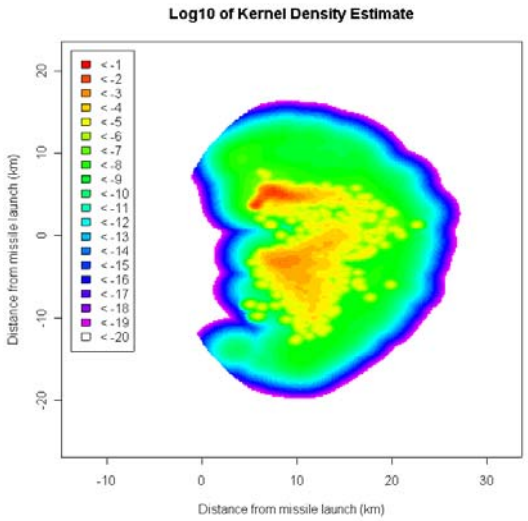

Figure 12: Overall MEBPDF for scenario in Table 2

2.4 Putting it all together

In this section we bring together the results developed in previous sections and provide algorithms for computing, for a given scenario:

- The expected number of casualties.
- An exclusion zone for a given exclusion probability level.
- The probability of a missile going outside a given area.

2.4.1 Predicting the number of people exposed to risk of injury

We are now in a position to estimate the expected number of people exposed to risk of injury for a given test scenario.

Denote by *POP* the matrix defining the population density over the same grid layout used to define PDF^S. The (x,y) co-ordinate of POP^S is the expected number of people in grid square *(x,y)*.

Let A be the relative size of the impact area of the missile compared to the size of a grid square.

Assuming that the impact point of a missile is within a grid square and that the impact area is always fully contained within that grid square, the expected number of people exposed to risk of injury, E[I], is given by

$$E[I] = A \sum_x \sum_y PDF^S(x,y) \,.\, POP(x,y), \quad or$$
$$E[I] = A \sum_x \sum_y MEBPDF^S(x,y) \,.\, POP(x,y),$$

depending on whether an MEB is available, and *A* is the fraction of the grid square that is affected when a missile impacts.

We noted earlier that it may not be necessary to accurately estimate the probability density function central to the impact envelope. When computing expected injury rates an accurate estimate for the entire region should be used. However, we have pragmatically assumed that a standard operational practice will be to clear people from the central impact region and hence any errors that might be introduced due to the less accurate interior probability density estimates will minimal.

Figure 13 shows an artificial population density map on the same grid as used to compute the probability density function developed in Section 2.3. As with the maximum energy boundary, the software used to compute the expected number of casualties allows the user to input an appropriate population density. It should be noted, however, that the input population density must be on the same grid as the kernel density estimate. Alternatively, it may be simpler to deliberately generate the kernel density estimate on the same grid as the population density function available.

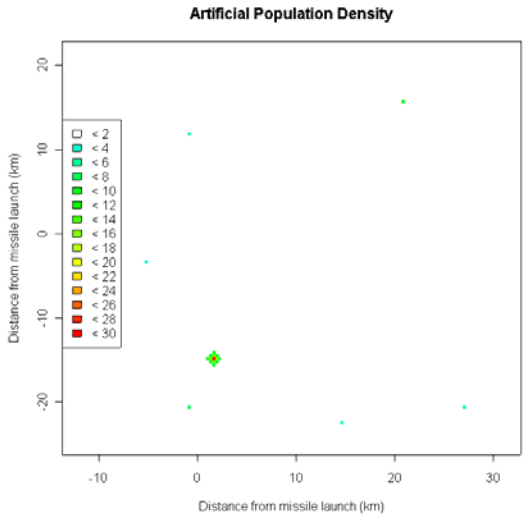

Figure 13: Artificial population density

With a nominal value of A of 0.1, in this example the expected number of casualties is 6.150934×10^{-10} if the PDF is used and is 6.117528×10^{-10} if MEBPDF is used.

2.4.2 Creating an exclusion zone

Of potential interest is the idea of determining a convex boundary around the missile range such that the probability of a missile impacting outside the boundary is less than some pre-determined level, for example, 10^{-6}. This boundary defines a potential "exclusion zone" for clearance of personnel and / or members of the public.

One method for creating such an exclusion zone with probability level ε is the following:

1. Sort *PDF*s from highest to lowest probability density across all grid squares.
2. Sum the sorted list from highest to lowest probability density, stopping when the total probability within the exclusion zone is greater than 1-ε.[1] Store the list of grid squares used in the sum.
3. Create a convex hull around the grid squares used in the sum of Step 2.

Figure 14 shows a convex 10^{-6} exclusion zone created around the probability density function developed in Section 2.3. The points plotted in Figure 14 represent the grid squares included in the raw exclusion zone. Figure 14 also shows the convex hull around this exclusion zone. Now, the raw exclusion zone, illustrated by the points shown in Figure 14, contains a total probability of at least $1 - 10^{-6}$. Therefore, the convex hull shown in Figure 14 is clearly a conservative 10^{-6} exclusion zone. It is possible to obtain a less conservative exclusion zone using, one of a number of different methods, but this would be considerably more complicated and more computationally intensive. Additionally, it may not be necessary in this

[1] Correct when the total probability across the PDF <= 1. Not correct when the PDF has been over estimated.

application, since the probability added by taking the convex hull of the raw exclusion zone may be very small. This is a possibility that may be worthy of further investigation.

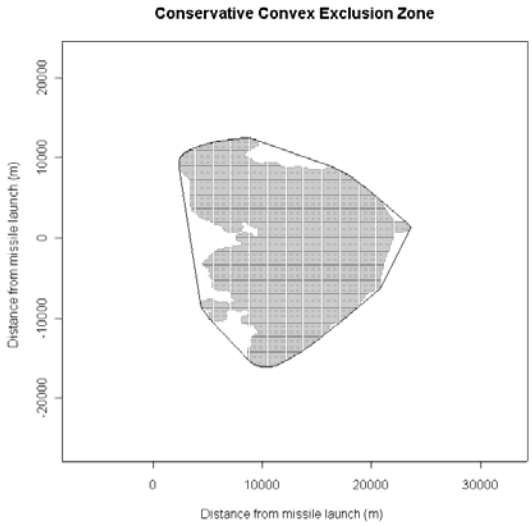

Figure 14: Convex 10^{-6} exclusion zone for scenario described in Table 2

2.4.3 Computing probability of leaving the range

Another possibility of potential interest is the idea of determining the probability that a missile leaves a firing range. This can be computed in a straightforward fashion from the data generated above.

Denote by *R* the indicator matrix defining the firing range over the same grid layout used to define PDF^S. The (x,y) co-ordinate of *R* is 0 if grid square *(x,y)* is within the firing range and 1 otherwise.

Then the total probability mass of PDF^S that falls outside the firing range, P_R , is given by

$$P_R = \Sigma_x \Sigma_y PDF^S(x,y) \,.\, R(x,y).$$

If the Maximum Energy Boundary is known, then P_R can be computed using

$$P_R = \Sigma_x \Sigma_y MEBPDF^S(x,y) \,.\, R(x,y).$$

Figure 15 shows a firing range boundary overlaid on the MEB probability density function developed in Section 2.3. In this example, the probability of the missile leaving the range is 1.954699×10^{-6}.

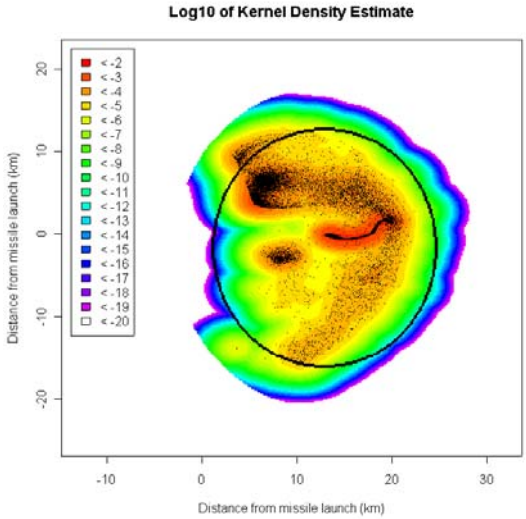

Figure 15: MEBPDF for scenario in Table 2 showing boundary of an artificial firing range

2.5 Symmetric Scenarios

Generating a database of impact-point data sets for all scenarios of interest can be an expensive operation. One idea that may make it simpler to populate such a database is the idea of symmetric scenarios, that is, whether symmetric initial conditions may produce symmetric impact densities. If this were the case, then it may be possible to obtain density estimates for multiple scenarios by computing a single kernel density estimate for a particular scenario and then using reflections of this density for other scenarios.

In order to test this hypothesis, kernel density estimates have been computed for two scenarios with symmetric initial conditions, and a measure of the difference between these kernel density estimates has been calculated, where one density was reflected. These kernel density estimates were based on samples of 20,000 observations. The measure of the difference between the densities is the same as that described in Section 2.2.3. If the hypothesis is true, then when one kernel density estimate is reflected, the kernel density estimates should be similar, and the measure of the difference between them should be small. In order to be able to assess the value of the difference that should be expected for two 'similar' densities, 20 pairs of disjoint samples from the same scenario were generated, and the differences between the kernel density estimates for each pair were calculated. These samples also contained 20,000 observations. From this sample of 20 differences, a mean, variance and 99% confidence interval have been calculated. The value of the difference between the kernel density estimates for the symmetric scenarios was then used to calculate a quantity known as the *p-value*. In this case, the interpretation of the p-value is that it gives the probability that the two kernel density estimates are exactly symmetric. If the p-value is large enough, then it may be reasonable to use reflected kernel density estimates to estimate the densities for symmetric scenarios. However, if the p-value is very small, then this technique may not be appropriate.

It must be noted, however, that this investigation has been carried out for only one pair of symmetric scenarios, so it may not be reasonable to extrapolate these results to apply to all scenarios. It must also be noted that the scenarios used in this investigation are symmetric in the x-axis. In order to be able to actually apply these ideas, it would be necessary to conduct a much more extensive investigation. An expert in the system under test, for example the senior simulation model engineer, may also be able to determine whether it is appropriate to apply these ideas to the ground impact data points based on their knowledge of the system's behaviour.

The following figures provide a visual indication of the extent of the symmetry of the impact distribution between the two scenarios. Figure 16 and Figure 17 show scatterplots for two symmetric scenarios, each with 20,000 observations. Since there were 50,000 observations available for the second symmetric scenario, Figure 17 simply shows a typical subset of 20,000 observations. Figure 18 shows a combined scatterplot of both scenarios, where the dataset corresponding to the second symmetric scenario has been reflected. We see in that figure that the "hook" regions align very closely, but visually there appears to be a significant difference between the two scenarios across the rest of the impact region. This suggests that either the data were generated differently for points outside of the "hook" or that the impact distributions are in fact different. Figure 19 and Figure 20 show kernel density estimates for the symmetric scenarios, where the density for the second scenario has again been reflected. The figures show that the ground impact points are not symmetric, despite the symmetry of the scenarios.

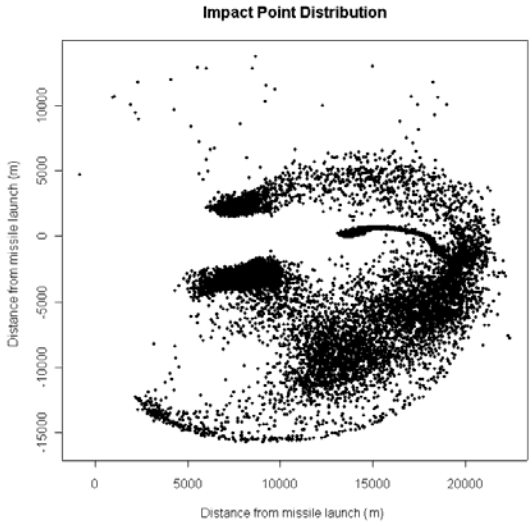

Figure 16: Scatterplot for first symmetric scenario

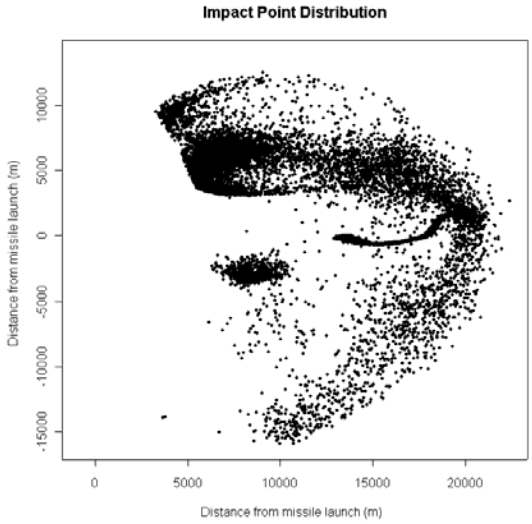

Figure 17: Scatterplot for second symmetric scenario

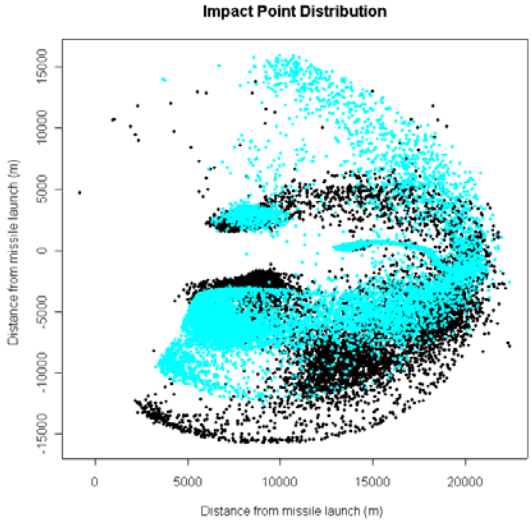

Figure 18: Scatterplot showing both symmetric scenarios (where one dataset is reflected)

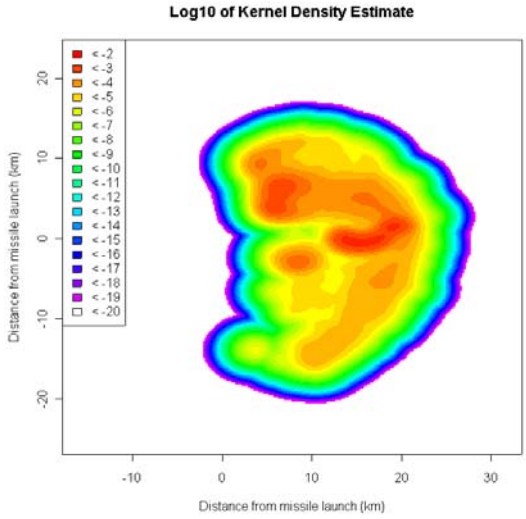

Figure 19: Kernel density estimate for first symmetric dataset

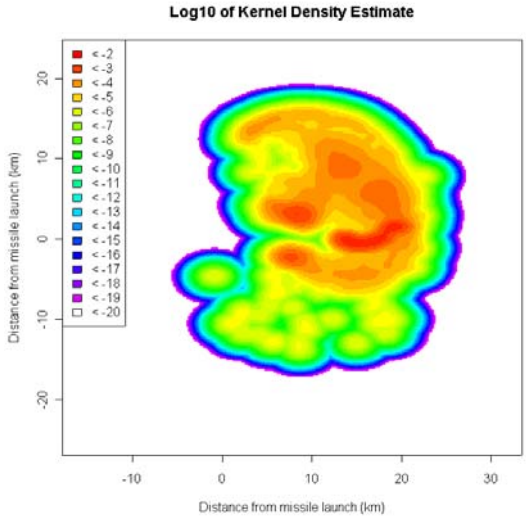

Figure 20: Reflected kernel density estimate for second symmetric dataset

3. Investigation of Appropriate Size of Datasets and Resolution of Kernel Density Estimates

3.1 Procedure of Investigation

The second set of tasks undertaken by CDCIN for the RSTT project included an investigation of the most appropriate number of observations and resolution of the KDEs (Kernel Density Estimates) to be generated.

Due to the amount of time required to generate each data point, and the vast number of datasets to be produced, it was deemed important to investigate the relationship between the number of observations used and the quality of the KDE produced.

A second factor affecting the quality of the KDE is the *resolution* of the KDE. Throughout this document, the term *resolution* will refer to the number of intervals into which the x- and y-dimensions of the impact distribution are divided. For example, for a KDE with a resolution of 16, the area of the impact distribution would be divided into a 16×16 rectangular grid, and the KDE process estimates the probability mass of the impact distribution that is contained within each rectangle. In order to determine the most appropriate number of observations, it was also necessary to investigate the relationship between the resolution and the accuracy of the KDE. We test the changing accuracy of a series of KDEs of a given scenario by examining the difference between an estimated KDE and the "overall KDE" of the impact distribution. It was also necessary to investigate the relationship between the number of observations and the resolution, since, for example, a 1000×1000 KDE based on only 10 observations would provide a misleading amount of detail, and this may affect the accuracy of the KDE.

In the following, we take as the overall KDE the average of 20 independent KDEs, based on 50,000 observations each. Therefore, the overall KDE is based on a total of 1,000,000 observations. The reason that we have used an average 20 KDEs based on 50,000 observations each, rather than a single KDE based on all 1,000,000 observations, is that computer memory constraints did not allow 1,000,000 observations to be processed simultaneously. The measures of difference will be referred to as the Mean Log Scaled (MLS) difference, the Exclusion Zone (EZ) difference, and the Total Log difference. Each of these functions implements a different way of measuring the difference between two probability mass functions defined over the same range, each highlighting slightly different aspects of this difference.

The data used in this investigation was supplied by DSTO from the data-set "fault-set all actuators zeroed GGM *012* 1 - 1,000,000", on 23 September 2005. All figures in this section, except for Figure 28, were generated by partitioning the data into subsets of a particular size, and calculating the average difference of the subsets from the overall KDE, for various sizes and resolutions. Therefore, all figures, except for Figure 28, are based on the entire dataset in the file mentioned above. Figure 28 is based on a particular subset of 100 observations from this file.

3.2 Results and Discussion

The first measure of difference calculated was the *Mean Log Scaled* (MLS) difference. The MLS difference for two KDEs was calculated by the following formula:

$$\frac{1}{n^2} \sum_{\text{all grid squares}} log_{10}\left(\frac{(p_{test} - p_{overall})^2}{p_{overall}(1 - p_{overall})} \right),$$

where p_{test} is the probability mass for the KDE being tested, for a particular grid square, $p_{overall}$ is the corresponding probability mass for the overall KDE, and n is the resolution of the KDEs.

That is, for each grid square, take the square of the difference between the two probability masses, divide by a scaled version of the 'correct' KDE, take the log (base 10). Finally, take the average of this quantity over all grid squares.

The scaling factor has been chosen such that an overall probability of p is treated symmetrically to an overall probability of 1-p., and the log has been taken in order to emphasise larger differences.

A typical plot of the average "error" of a test KDE from the overall KDE against the resolution, for a fixed number of observations is shown in Figure 21.

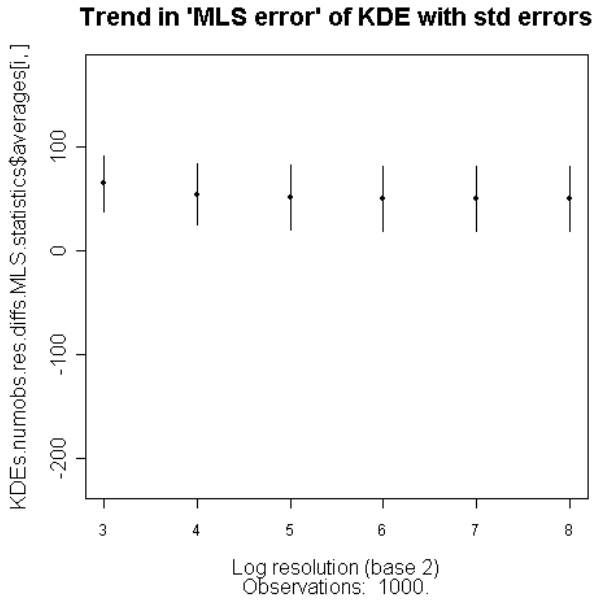

Figure 21: Average Mean Log Scaled error vs resolution for 1000 observations

Figure 21 indicates that as the resolution is increased, the "error" of the KDE initially decreases by a small amount, but for a resolution greater than 32 (5 on the x-axis of the figure as it is shown as the log to base 2), the improvement is negligible. A second point illustrated by Figure 21 is that as the resolution is increased, the standard deviation of the "errors"

initially increases by a small amount. However, again, for a resolution greater than 32, the difference in standard deviation is negligible. For these reasons, it was decided that resolutions of 16 and 32 were of greatest interest, and for the remaining measures of difference, plots were only generated for the average error against the number of observations, for fixed resolutions of 16 and 32. These plots are shown in Figure 22 - Figure 26.

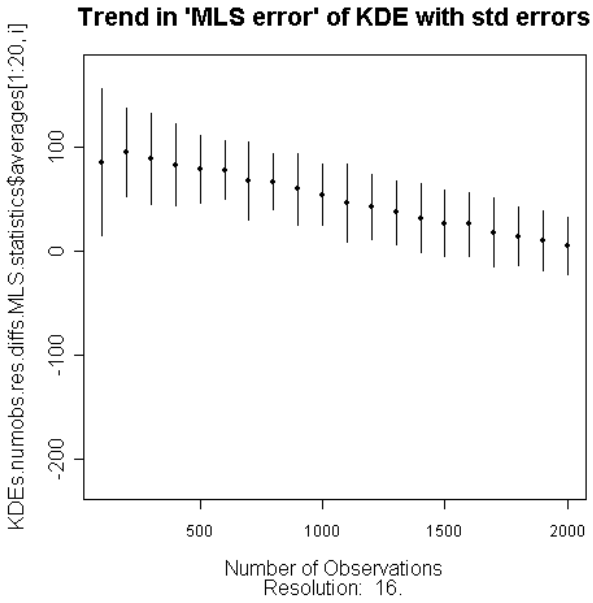

Figure 22: Average Mean Log Scaled error vs number of observations for a resolution of 16

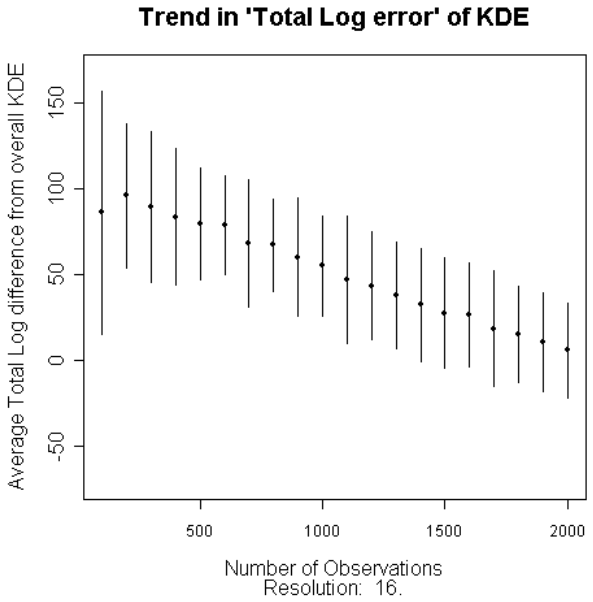

Figure 23: Average Total Log error vs number of observations for a resolution of 16

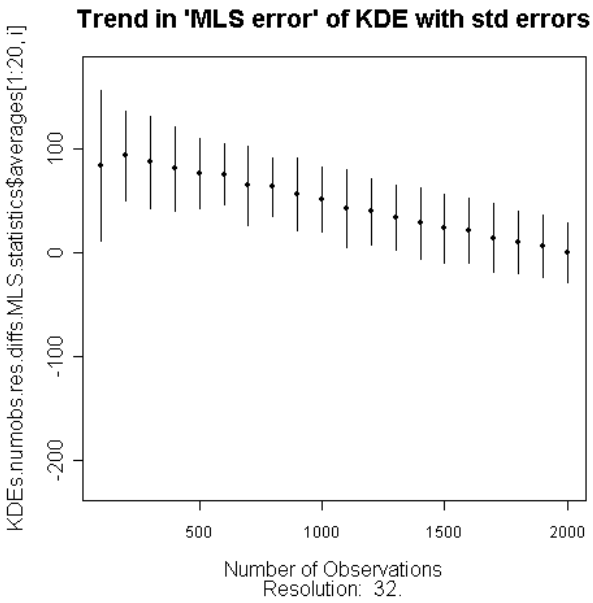

Figure 24: Average Mean Log Scaled error vs number of observations for a resolution of 32

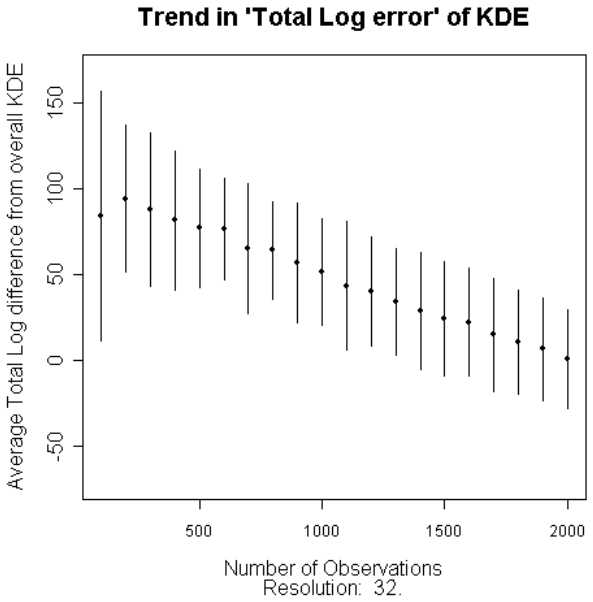

Figure 25: Average Total Log error vs number of observations for a resolution of 32

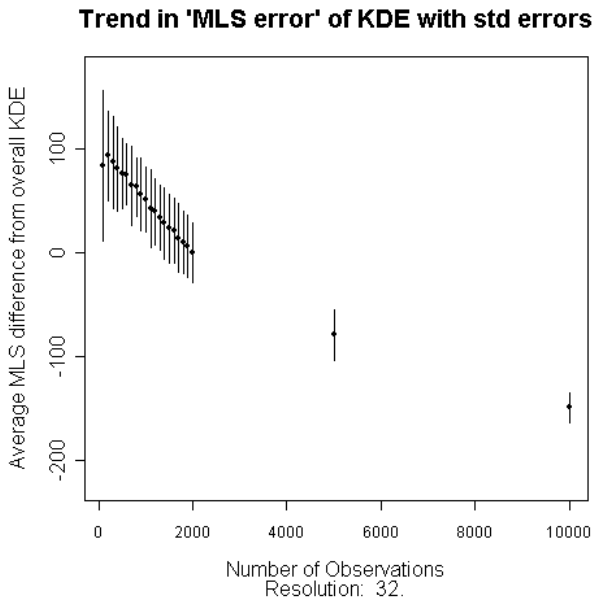

Figure 26: Average Mean Log Scaled error vs number of observations for a resolution of 32

The most obvious observation from Figure 22 - Figure 26 is that as the number of observations increases, it appears that the "error" of the KDE from the overall KDE will continue to decrease indefinitely. Therefore, in the feasible range of the number of observations, there appears to be no threshold above which a "perfect" KDE is obtained, for which no further improvement can be gained. Similar analysis during Stage 1 of the RSTT project also suggested that no such threshold exists up to 35,000 observations.

The second point to observe from Figure 22 - Figure 26 is that as the number of observations increases, the standard error also appears to decrease. As with the average error, this trend also appears to continue indefinitely, as can be observed from the standard error bars for 5000 and 10,000 observations, in Figure 26, above.

The final measure of difference calculated was the exclusion zone difference. Figure 30 and Figure 31, below, show the average errors with standard error bars, for the Exclusion Zone difference. The Exclusion Zone difference is made up of two components. These components are plotted in green and red in Figure 30 and Figure 31, below. The green components represent the percentage of grid squares that were conservatively included in the exclusion zone for the test KDE and the red components represent the percentage of grid squares that were incorrectly omitted from the exclusion zone for the test KDE. The black components represent the total Exclusion Zone error, which is the sum of the red and green components. Note that each of these figures is given as a percentage of all grid squares included in either the given test KDE, the overall KDE or both.

Figure 27 and Figure 28, below, show exclusion zones for the overall KDE and a test KDE based on 100 observations, with a resolution of 16. Figure 29 shows the components of the exclusion zone difference for these two exclusion zones. For this particular test KDE, the percentage of grid squares that were conservatively included in the exclusion zone for the test

KDE (that is, the green component) is equal to 14.06%, and the percentage of grid squares that were incorrectly omitted from the overall KDE (that is, the red component) is equal to 16.68%. Therefore, the total exclusion zone error between the two exclusion zones shown in Figure 27 and Figure 28 is equal to 30.74%.

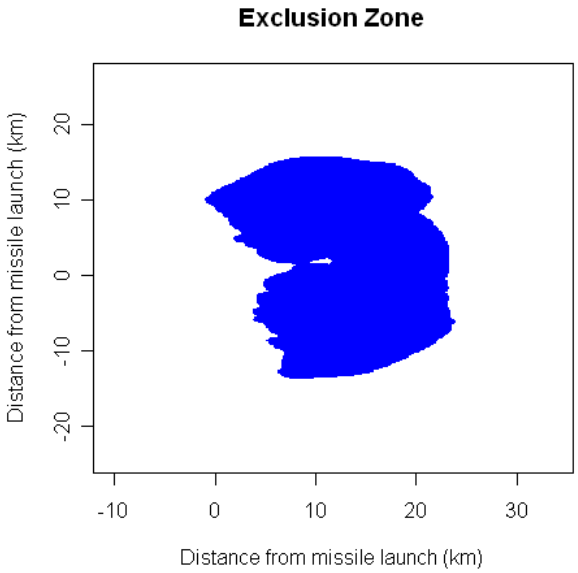

Figure 27: Overall exclusion zone

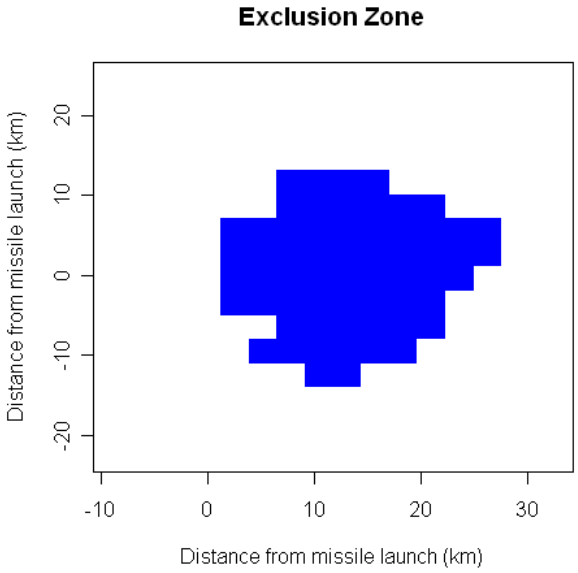

Figure 28: Exclusion zone for a KDE with 100 observations and a resolution of 16

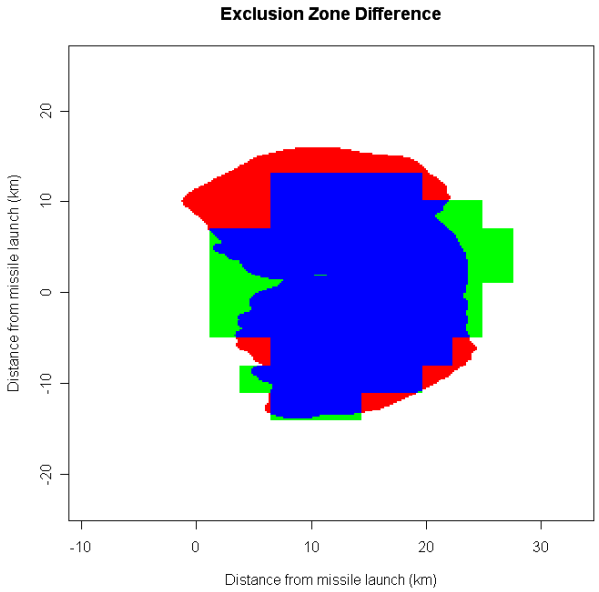

Figure 29: Components of the exclusion zone difference for the two exclusion zones shown above

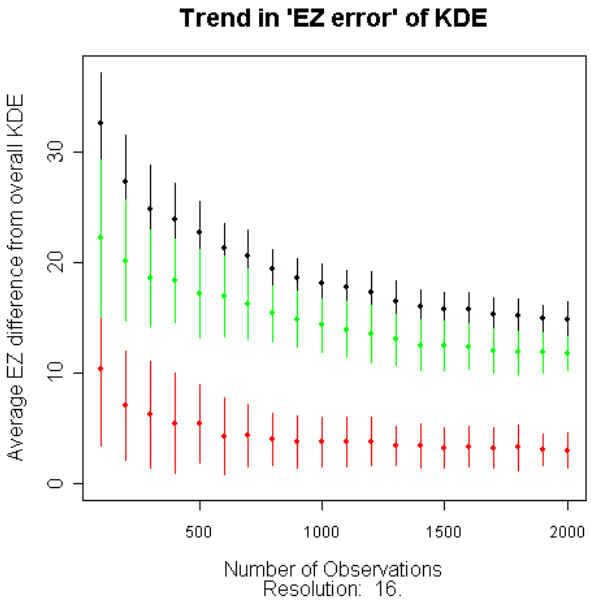

Figure 30: Average Exclusion Zone error vs number of observations for a resolution of 16

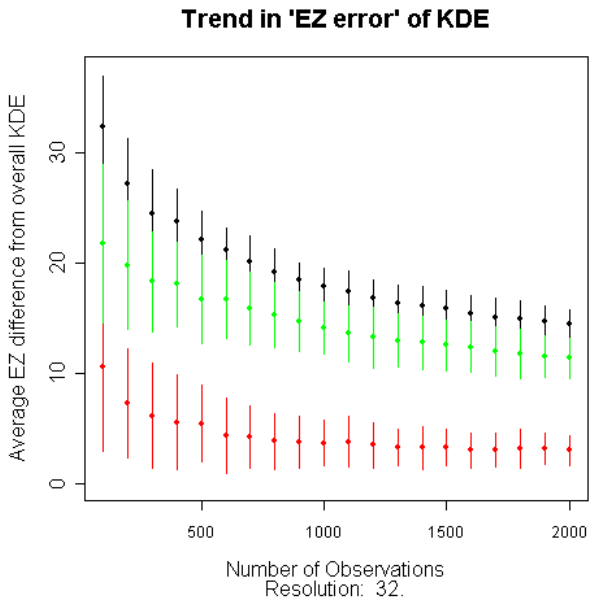

Figure 31: Average Exclusion Zone error vs number of observations for a resolution of 32

In contrast with the earlier figures, Figure 30 and Figure 31 indicate that the improvement in the accuracy of the exclusion zone diminishes as the number of observations increases.

The Exclusion Zone difference also appears to have a number of other useful qualities. It seems to have the smallest standard error of all the measures of difference calculated, and it also seems to give the most consistent and smooth trend as the number of observations increases. Additionally, it can be seen that the Exclusion Zone error is composed predominantly of the conservative portion of the error, plotted in green.

Regarding the selection of the number of observations, the red parts of Figure 30 and Figure 31 suggest that if the number of observations is at least 600, then the average percentage of grid squares incorrectly omitted from the exclusion zone is around 5%. It is also likely that this percentage will be considerably reduced when the exclusion zone is converted to a convex hull. With the current software, it is not possible to determine how much this percentage would be reduced, because it is not possible to apply this measure of difference to the convex version of the exclusion zone.

Overall, it appears that the quality of the KDE generated continues to improve as the number of observations is increased. This suggests that as many observations as possible should be generated, since any increase will result in improved estimates of the KDE. However, the observed trends in the Exclusion Zone difference measure suggest that if any less than 600 observations are generated for each scenario, the percentage of grid rectangles incorrectly omitted from the exclusion may be unnecessarily high, of the order of 5%. Consequently, 600 observations is recommended as a lower bound on the number of observations generated for a given scenario.

3.3 Recommendations

We recommend the following:

1. That KDE resolutions beyond 16 x 16 and 32 x 32 do not provide significantly more accurate information and hence 16 x 16 or 32 x 32 resolutions appear to be suitable for the development of Range Safety Templates.
2. That at least 600 observations (impact data points) be used in generating KDEs for a given scenario.
3. That a more precise estimate of the average percentage of grid squares incorrectly omitted from the exclusion zone be obtained.

4. Issues in KDE generation – Dynamic Bandwidth Selection

The third set of tasks undertaken by CDCIN for the RSTT project included researching problems adjacent to the PDF generation problem. During the course of this third set of tasks however, it became apparent that for input datasets with certain properties, the kernel density estimation algorithm proposed previously produced an inappropriate kernel density estimate. An example of such a KDE is given in Figure 32, below.

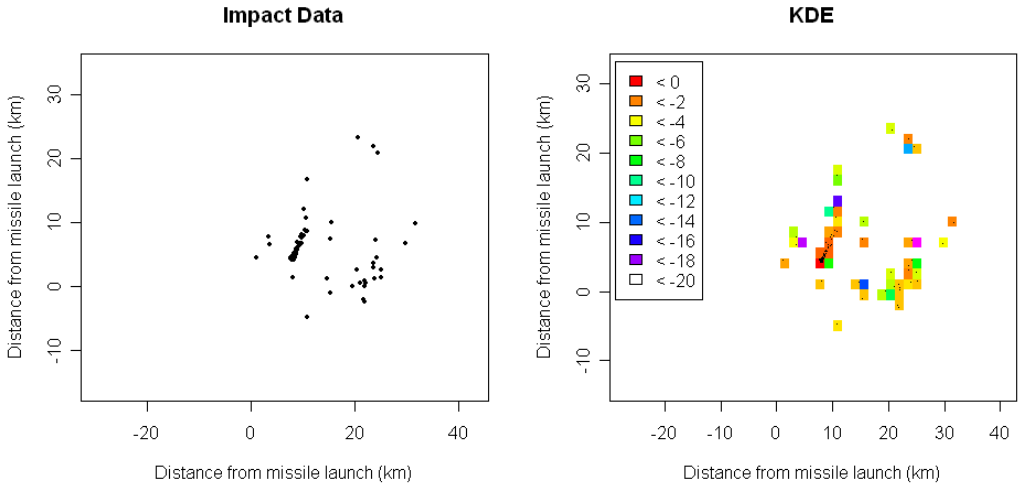

Figure 32: Scatterplot and KDE with inappropriate bandwidth selection

The impact data comprises 600 points, with the bulk of the points (approximately 90%) concentrated around the point (8,4).

Ideally, the KDE for this dataset should contain reasonably high levels of probability for the majority of the area within the convex hull of the impact points. As can be seen in the KDE in Figure 32, there are large areas within the convex hull that contain effectively 0 probability mass. Upon further investigation, it was found that the source of this problem was the bandwidth selection algorithm. More specifically, it was found that if the dataset input to the KDE algorithm contained a very dense cluster of points, the bandwidths selected by the algorithm were too small, and the resulting KDE exhibited features such as those seen in Figure 32.

A possible solution to this problem would be to generate separate KDEs for the cluster and the remaining points, then taking a weighted combination of the two KDEs. This can be achieved by the procedure:

1. Determine a grid over which the final KDE is to be generated.
2. Identify a small set of grid squares containing high densities of points.
3. Generate a KDE for the points within the small set of grid squares.
4. Generate a separate KDE for the remaining points.
5. Take a weighted combination of the two KDEs.

Note that a KDE is represented by a matrix where each cell of the matrix contains the estimated probability mass for a grid square Therefore, the weighted combination referred to in Step 5 of the above procedure can be calculated by the following formula:

$$KDE_{final} = p \times KDE_{cluster} + (1-p) \times KDE_{scatter}$$

where $KDE_{cluster}$ is the KDE based on the points in the chosen set of dense grid squares, $KDE_{scatter}$ is the KDE based on the points in the remaining points, scattered outside the chosen set of grid squares, p is the proportion of impact points that lie within the chosen set of grid squares.

Figure 33, below, shows the KDE generated by this method for the scenario shown in Figure 32. Note that as this method uses different bandwidths for different sets of points, this is actually an example of a KDE method using dynamically determined bandwidths.

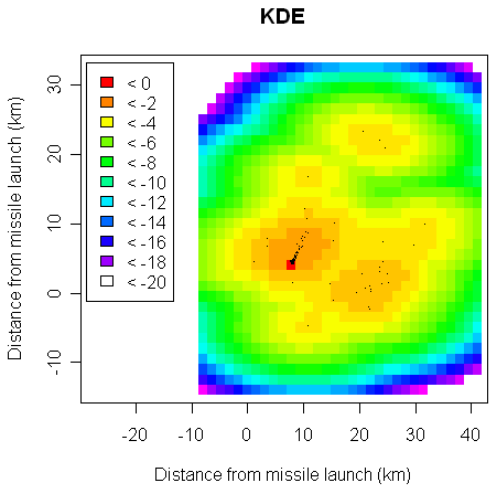

Figure 33: KDE with dynamic bandwidth selection

In order to develop a more robust KDE algorithm, it is necessary to ensure the selection of an appropriate bandwidth, preferably via an automated procedure.

We commence by examining the effect on KDE generation of the existence of a tight cluster of points in the data set. We have created data sets in which between 10% and 90% of the points are specifically allocated within one grid square and the remainder are uniformly distributed over a larger region. We then generate KDEs for each case and identify where the KDEs become inappropriate. Figure 34 to Figure 42 show the difference between KDEs generated using the "standard" static bandwidth procedure and using the procedure outlined in Steps 1-5, above.

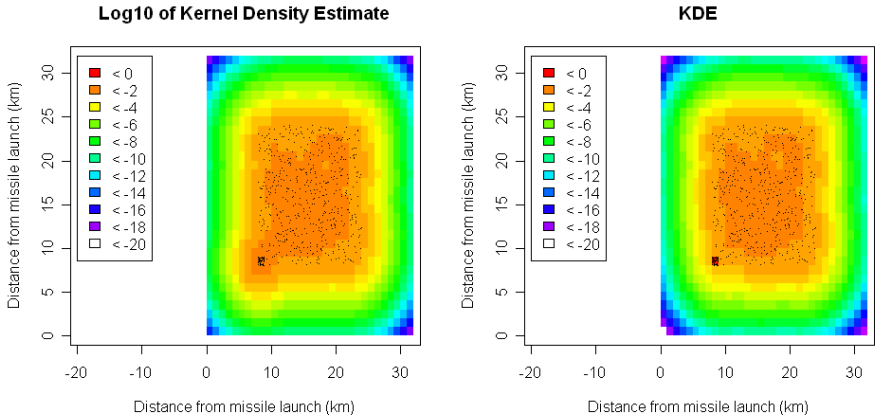

Figure 34: Static, on the left, and dynamic, on the right, bandwidth KDEs with a 10% cluster. Note that both plots are log base-10 of the estimated probability density.

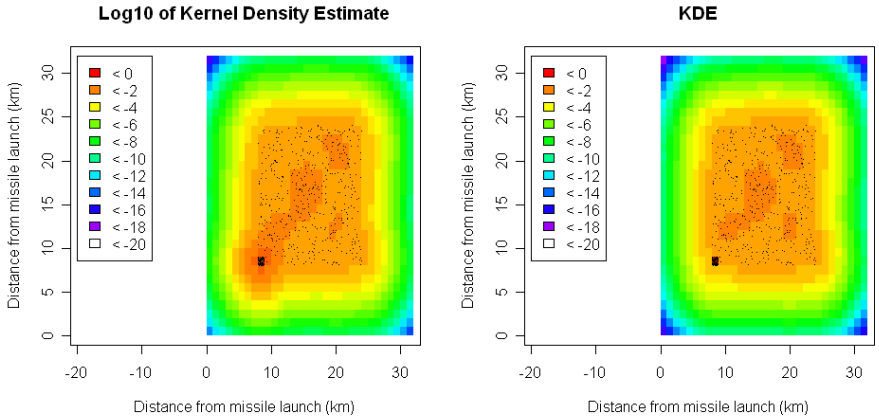

Figure 35: Static and dynamic bandwidth KDEs with a 20% cluster. Note that both plots are log base-10 of the estimated probability density.

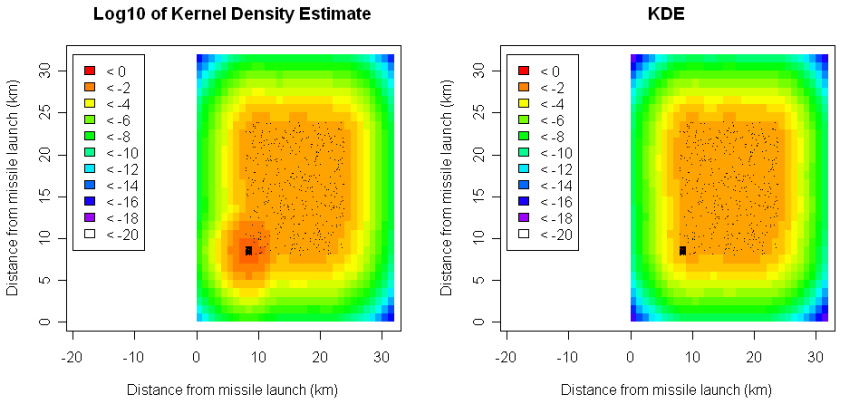

Figure 36: Static and dynamic bandwidth KDEs with a 30% cluster. Note that both plots are log base-10 of the estimated probability density.

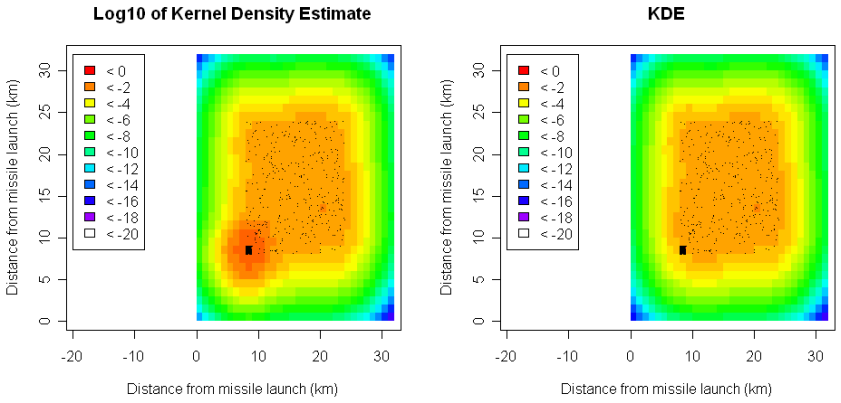

Figure 37: Static and dynamic bandwidth KDEs with a 40% cluster. Note that both plots are log base-10 of the estimated probability density.

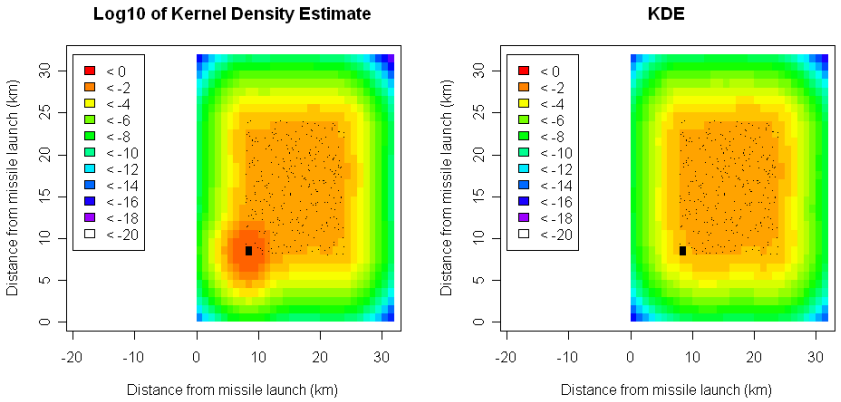

Figure 38: Static and dynamic bandwidth KDEs with a 50% cluster. Note that both plots are log base-10 of the estimated probability density.

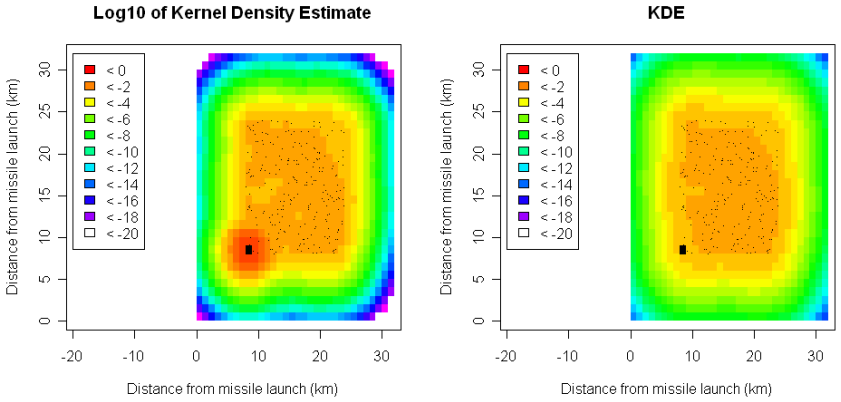

Figure 39: Static and dynamic bandwidth KDEs with a 60% cluster. Note that both plots are log base-10 of the estimated probability density.

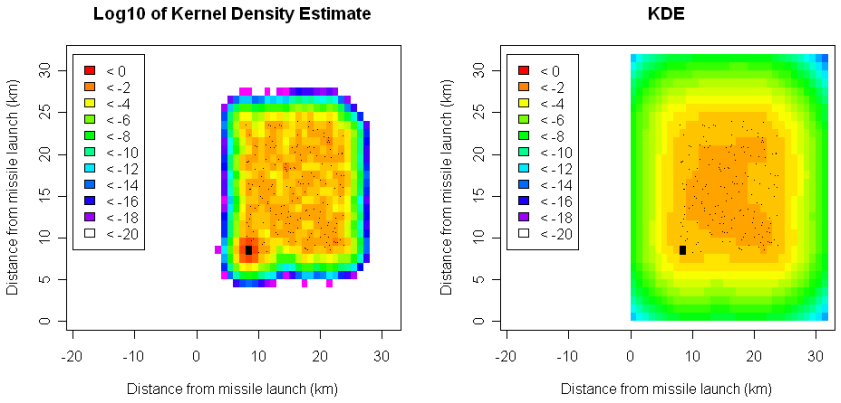

Figure 40: Static and dynamic bandwidth KDEs with a 70% cluster. Note that both plots are log base-10 of the estimated probability density.

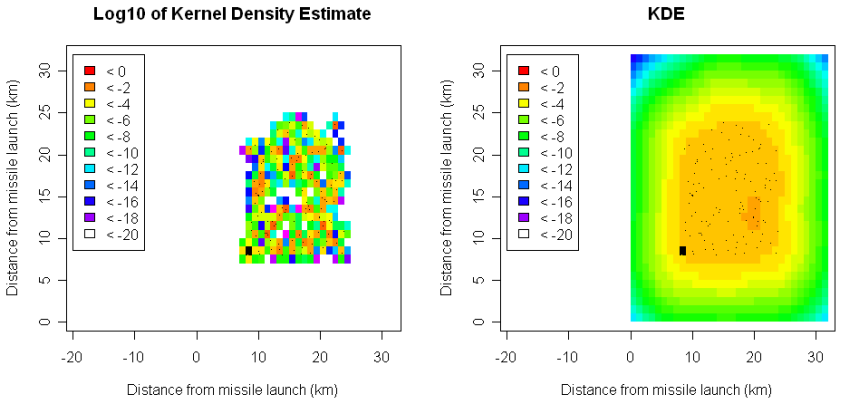

Figure 41: Static and dynamic bandwidth KDEs with a 80% cluster. Note that both plots are log base-10 of the estimated probability density.

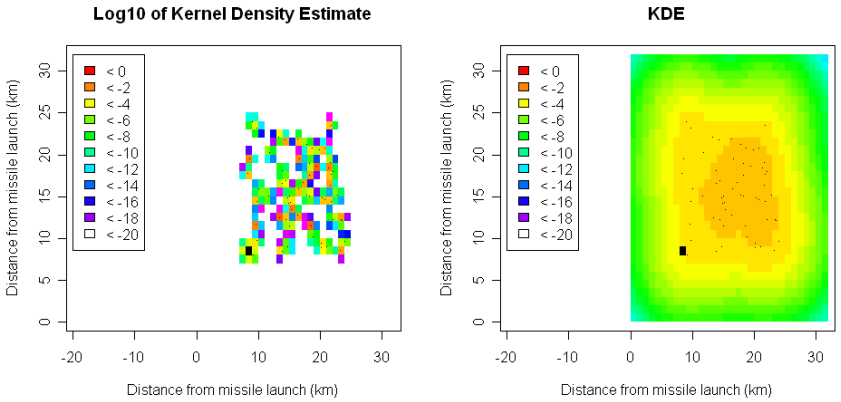

Figure 42: Static and dynamic bandwidth KDEs with a 90% cluster. Note that both plots are log base-10 of the estimated probability density.

Note that in all of the plots shown in Figure 34 to Figure 42, the colour of the grid square in which the cluster occurs is somewhat obscured by the impact points plotted in this grid square. However, in each of these plots, the grid square in which the cluster occurs will contain a large probability mass, and would therefore appear as a red grid square in these plots, if it were not obscured.

For the scenarios shown in Figure 34 to Figure 42, with up to 50% of impact points in a single grid square, the statically generated KDE and dynamically generated KDE are very similar. However, with 70% in a single grid square, the effective ranges of the two KDEs are very different, while at 80% there are patches of effectively 0 probability within the convex hull of the data. This suggests that a dynamic bandwidth KDE generation procedure that searches for cluster densities of, say, 30% of impact points will provide very similar KDEs for lower cluster densities, and effectively "spread" probability mass over the wider impact area in the presence of higher impact densities.

We note that Figure 34 to Figure 42 also highlight a potential limitation of the proposed dynamic bandwidth procedure. The KDEs generated by the dynamic bandwidth procedure suggest an extremely tight boundary around the cluster area, and the area immediately outside the impact area has a reasonably low impact probability. Whilst this may be justified based on examination of the numerical data purely in isolation, care must be taken in the actual application of the results given potential uncertainties in the physical nature of the process being modelled and the method by which the data is being generated (numerical simulation).

We recommend that relevant subject matter experts independently review the KDEs generated in such cases and provide advice on the subsequent construction of range safety templates, in particular near areas of dense concentration of impact points.

We recommend that if 30% or more of impact points lie within a single grid square that the dynamic bandwidth KDE procedure described above be employed.

We also note that the issue identified when points are tightly clustered also occurs for datasets containing a dense, narrow band of points, as illustrated in the following example.

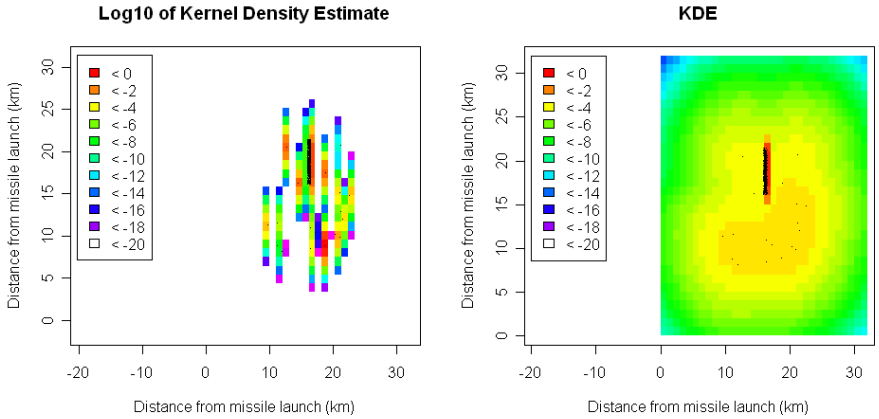

Figure 43: Static and dynamic bandwidth KDEs with a concentrated band of impact points. Note that both plots are log base-10 of the estimated probability density.

Figure 43 shows KDEs resulting from a dataset containing a dense narrow band of points. The static bandwidth KDE demonstrates a typical KDE based on a dataset containing a dense narrow band of points, and the dynamic bandwidth KDE demonstrates how the KDE can be modified by applying a variant of the dynamic bandwidth procedure (details provided below).

Finally, if a dataset contains multiple clusters or bands of points, the bandwidth selection issues described above do not occur, provided that the clusters or bands are sufficiently spaced apart. This is demonstrated by Figure 44, below.

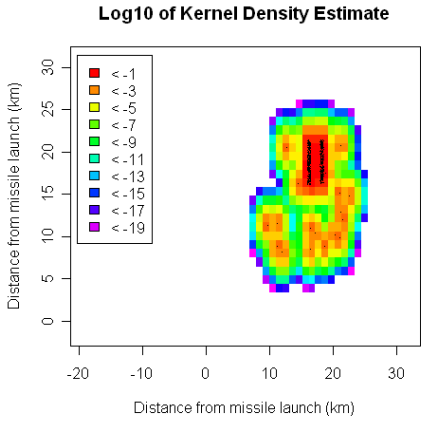

Figure 44: Static bandwidth KDE with a two concentrated bands of impact points

The figure shows two high density impact bands, in adjacent “columns” of grid squares. There is sufficient variation in the density of impact points in the x-axis to ensure a reasonable selection of the bandwidth parameter, resulting in a reasonable spread of the probability mass over the entire impact area.

As a consequence of the above investigation, a potential process for dynamic bandwidth selection that addresses the above situation is as follows:

1. Determine the percentage of points in each individual grid square.
2. If any of these percentages exceed a certain threshold, then separate the points in the densest grid square from the dataset and take a weighted combination of the KDEs for the dense grid square and the remaining points.
3. If none of the percentages in Step 1 exceed the threshold, then determine the percentage of points in each row and column of grid squares.
4. If any of the percentages in Step 3 exceed a certain threshold, then separate the points in the most dense row or column from the dataset and take a weighted combination of the KDEs for the dense row or column and the remaining points.
5. If none of the percentages in Step 3 exceed the threshold, then generate a KDE using the static bandwidth method developed previously in the RSTT project.

5. KDE Isotropy

5.1 Observations from impact data

Further investigation of the growing collection of RSTT impact data sets has revealed that sub-optimal kernel density estimates are produced in some interesting cases. As shown in Figure 45, the two dimensional formula presented in section 2.2.1 produces a good kernel density estimate for data randomly scattered about the Y axis. In this case, the X and Y bandwidths are treated independently and correspond to the diagonal bandwidth matrix:

$$\mathbf{H} = \begin{pmatrix} h_x & 0 \\ 0 & h_y \end{pmatrix}$$

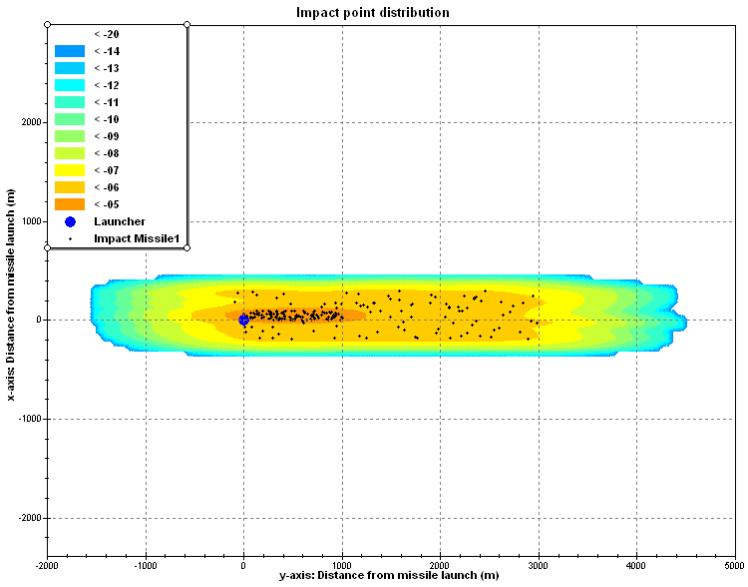

Figure 45: Diagonal bandwidth matrix KDE for impacts randomly distributed about the X and Y axis

However, in a number of the data sets examined the ground impacts appeared to be randomly distributed about axes rotated with respect to the X and Y axes. A representative ground impact distribution is shown in Figure 46. When the diagonal bandwidth matrix is applied in this case, the kernel density estimate is not as optimal as the example shown in Figure 45. For the purposes of this discussion, the ground impacts shown in Figure 46 are simply a forty-five degree rotation of the impacts presented in Figure 45. The KDE in Figure 46 is a less conservative result than the KDE shown in Figure 45 as several impact points now lie in the 1×10^{-7} probability region.

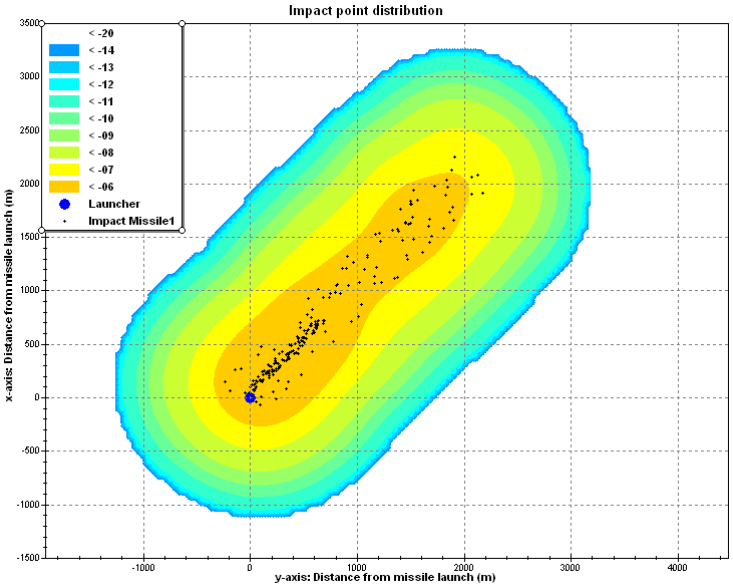

Figure 46: Diagonal bandwidth matrix KDE for example ground impact distribution

The discovery of cases that are clearly better represented by bandwidths defined about axes rotated with respect to X and Y has motivated some investigation into using a non-diagonal bandwidth matrix:

$$\mathbf{H} = \begin{pmatrix} h_x & h_{xy} \\ h_{yx} & h_y \end{pmatrix}$$

Defining the terms of the covariance matrix is not a trivial task and has received limited treatment in the literature. Recent work by Zhang et al. [5, 6] outlines Markov Chain Monte Carlo algorithms for estimating the bandwidth matrix parameters. Duong et al. [7, 8] discuss the application of plug-in algorithms, biased cross-validation and smooth cross-validation algorithms for defining the bandwidth matrix. With all proposed techniques the observed performance must be balanced against the computational effort required for large data sets. Based on the observed cases illustrated here, we propose one approach for deriving the terms of the full bandwidth matrix for guided weapon ground impact data.

5.2 Impact coordinate correlation

The two dimensional formula for $\hat{f}(x, y)$ presented in the section 2.2.1 is most applicable in cases where X_i and Y_i are sampled independently. In reality, the coordinates of the impact are not independent. The correlation between the X and Y coordinates of the sampled data is included in the kernel density estimate as follows:

$$\hat{f}(x, y) = \frac{1}{n}\sum_{i=1}^{n} K\left(x - X_i, y - Y_i, \mathbf{H}\right),$$

where H is the full bandwidth matrix:

$$\mathbf{H} = \begin{pmatrix} h_x & h_{xy} \\ h_{yx} & h_y \end{pmatrix}$$

The kernel, as represented in the above equation, is now two-dimensional and dependent on the bandwidth operator **H**. Ideally, **H** should not be degenerate, which will be the case for real 2D impact data. However, in cases where the determinant is small, the resolution error must be used to set a minimum value of the bandwidth elements. Note that this problem exists regardless of what formula is being used.

If the normal distribution represents the kernel, the estimator is the sum of the bivariate normal distributions:

$$\hat{f}(x,y) = \frac{1}{2\pi n\sqrt{\det(\mathbf{H}_2)}} \sum_{i=1}^{n} \exp\left(-\frac{1}{2}(x - X_i, y - Y_i)\mathbf{H}_2^{-1}(x - X_i, y - Y_i)^T \right)$$

The matrix $\mathbf{H}_2$ is the equivalent of the covariance matrix expressed in terms of bandwidth matrix **H**. The elements of the matrix $\mathbf{H}_2$ can be found using the one-dimensional formula along the *principal axes* of the distribution and then rotating the matrix back to the original coordinates. The principal axes of the distribution are the coordinates in which the XY-correlation is zero and correspond to the eigenvectors of the covariance matrix **S**:

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} \begin{pmatrix} (X_i - \bar{X})^2 & (X_i - \bar{X})(Y_i - \bar{Y}) \\ (X_i - \bar{X})(Y_i - \bar{Y}) & (Y_i - \bar{Y})^2 \end{pmatrix}$$

The covariance matrix represented in principal axes coordinates is a diagonal matrix consisting of h_x and h_y from the XY-independent formula. Suppose that the diagonal terms are S_1 and S_2 (eigenvalues) and the main eigenvector is (s_x, s_y), then:

$$\begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} = \mathbf{U}^{-1}\mathbf{S}\,\mathbf{U}$$

where

$$\mathbf{U} = \begin{pmatrix} s_x & -s_y \\ s_y & s_x \end{pmatrix}$$

U is the rotation matrix to the principal axes with the properties $|\mathbf{U}| = 1$ and $\mathbf{U}^{-1} = \mathbf{U}(s_y \to -s_y)$. The bandwidth calculated along the principal axes can be translated back to the original coordinate system using:

$$\mathbf{H}_2 = \mathbf{U}\begin{pmatrix} h_1^2 & 0 \\ 0 & h_2^2 \end{pmatrix}\mathbf{U}^{-1}$$

Note that only matrices $\boldsymbol{S}$ and $\mathbf{H}_2$ are rotated, not the coordinates of the impact points. The elements h_1 and h_2 are calculated as per the one-dimensional case using S_1 and S_2 values. If the value S_2 is small it should be increased to the resolution error factor (S_1 cannot be small since it is the main eigenvalue). A small S_2 value corresponds to the case where all points lie on a line or are very close to one.

If the formula from the section 2.2.3.3 is used for calculating bandwidth, then IQR should be calculated along the principal axes of $\boldsymbol{S}$. This is done by projecting each point onto the principal axes via:

$$(P_i, Q_i) = (X_i, Y_i)\,\mathbf{U}$$

The new P_i coordinates are then used to calculate h_1 and Q_i coordinates to calculate h_2. Figure 47 shows the KDE produced using the non-diagonal bandwidth matrix for the example data first presented in Figure 46. The KDE appears to be more consistent with the expected result demonstrated in Figure 45 for the non-rotated ground impact data.

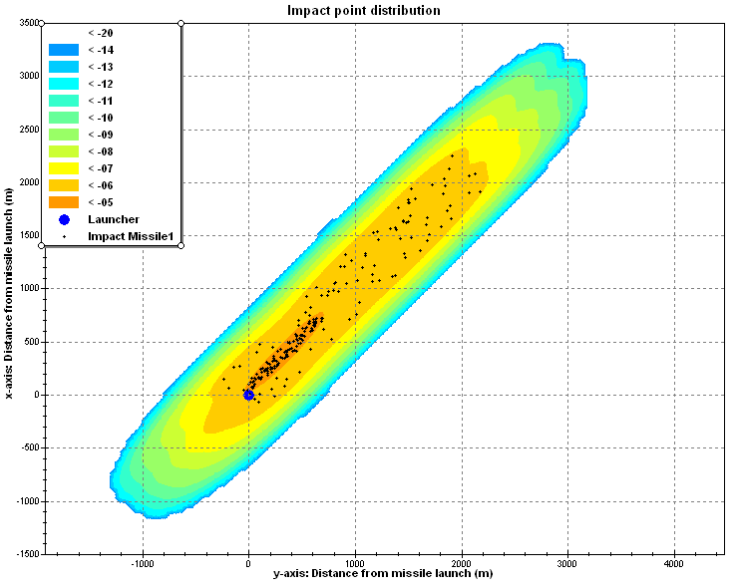

Figure 47: Full bandwidth matrix KDE for example ground impacts

It is possible to predict a distribution that will not be well represented by a KDE generated using the non-diagonal bandwidth matrix outlined here, as shown in Figure 48. This example is a hypothetical case that has not been observed in the available data sets. Figure 48 shows again that clustering (see section 4) can cause the global approach to bandwidth estimation to produce a non-optimal KDE in some cases. To date, there have not been any identified cases in the data that result in a less optimal KDE when using the non-diagonal bandwidth matrix as compared to using a diagonal bandwidth matrix.

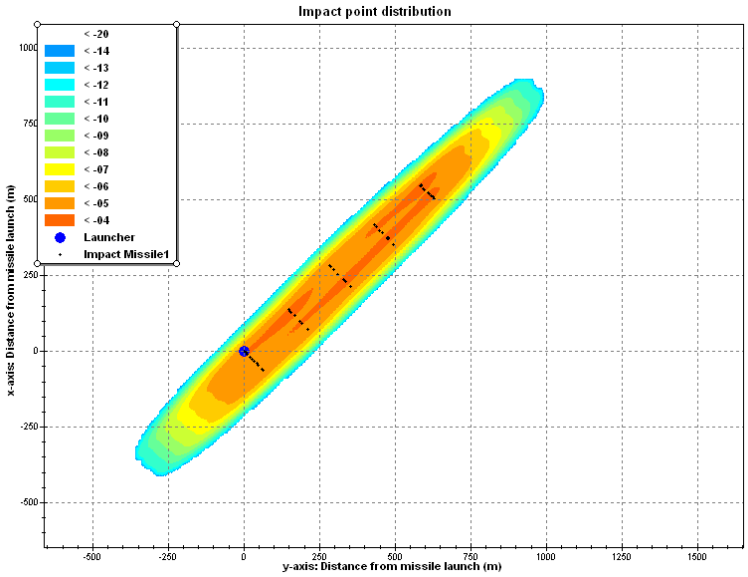

Figure 48: Data set with non-optimal KDE using a non-diagonal bandwidth matrix

5.3 Computational efficiency and accuracy

Calculating the grid values of the estimator $\hat{f}(x,y)$ using the above formula requires $O(nM^2)$ exponents, where M is the grid size and n is number of impacts. It is possible, however, to reduce the number of calculations by factorising the exponent in the bivariate normal distribution. Let us denote the coordinates of k-th impact as (X_k, Y_k). The bivariate normal distribution can then be written in a well known covariance form as:

$$\hat{f}(x,y) = \frac{\alpha}{n}\sum_{k=1}^{n}\exp(\beta\, z_k)$$

where

$$\alpha \equiv \frac{1}{2\pi\ \sigma_x\sigma_y\sqrt{1-\rho^2}},$$

$$\beta \equiv -\frac{1}{2(1-\rho^2)},$$

$$z_k \equiv \frac{(x-X_k)^2}{\sigma_x^2} + \frac{(y-Y_k)^2}{\sigma_y^2} - \frac{2\rho(x-X_k)(y-Y_k)}{\sigma_x\sigma_y},$$

and ρ is defined from $\mathbf{S}$

$$\mathbf{S} = \begin{pmatrix} \sigma_x^2 & \sigma_x\sigma_y\rho \\ \sigma_x\sigma_y\rho & \sigma_y^2 \end{pmatrix}$$

The expression can be factorised into the following form:

$$\hat{f}(x_i, y_j) = \hat{f}_{ij} = \frac{\alpha}{n} \sum_{k=1}^{n} B_{ij}^{(1)} B_{ik}^{(2)} B_{jk}^{(3)}$$

with

$$B_{ij}^{(1)} = \exp\left(-\beta \frac{2\rho x_i y_j}{\sigma_x \sigma_y} \right)$$

$$B_{ik}^{(2)} = \exp\left(\beta \frac{(x_i - X_k)^2}{\sigma_x^2} + \beta \frac{2\rho x_i Y_k}{\sigma_x \sigma_y} \right)$$

$$B_{jk}^{(3)} = \exp\left(\beta \frac{(y_j - Y_k)^2}{\sigma_y^2} + \beta \frac{2\rho X_k (y_j - Y_k)}{\sigma_x \sigma_y} \right)$$

This form is equivalent to the original covariance form (which can be verified by direct substitution), but now matrices $\mathbf{B}^{(2)}$ and $\mathbf{B}^{(3)}$ only need to be calculated once for all combinations *i*, *j*, and *k*. This will require about $O(2nM)$ exponent evaluations. This symmetry is a consequence of the grid being presented as a lattice aligned along X and Y axes and the bivariate normal distribution being symmetrical relative to its own principal axis. Using the **B** matrices, the exponents can now be calculated for $O(M^2+2nM)$ terms, which is less than the $O(nM^2)$ terms required previously.

There is a numerical side effect to this approach due to limitations in machine precision. In the original covariance form of the estimator, the exponent is always negative. While it can have a large absolute value when the test point is far from the impact point under consideration, the negative value means it has minimal contribution to the sum and so does not introduce significant numerical error. When using the estimator form containing **B** matrices, the exponent can be arbitrarily big: either negative or positive. Evaluation of the exponential can magnify the numerical error, which does not cancel out after matrix multiplication even if the result is a small value. This makes direct calculations, i.e. calculating exponents and then multiplying their values, undesirable.

One possible fix to this problem involves re-defining the exponents in the following way:

$$\ln B = p + aq$$

where *a* is an arbitrarily selected parameter, *q* is some integer number and $0 \le p < a$ is the remainder. The product of the exponential terms can then be calculated by multiplying *p* terms and summing *q* terms as the estimator now takes the form:

$$\hat{f}(x_i, y_j) = \hat{f}_{ij} = \frac{\alpha}{n} \sum_{k=1}^{n} e^{p_{ij} + aq_{ij}} e^{p_{ik} + aq_{ik}} e^{p_{jk} + aq_{jk}}$$

The parameter *a* is selected so that finding *p* and *q* is fast. After multiplying the *p* terms, the *q* terms are summed to give the factor $\exp(a)^{\Sigma q}$. All terms of this form can be obtained from a

pre-calculated table. Theoretically, the sum of the q terms is not bound by any limits. However, if the extreme lower value of the estimator evaluation is expected to be around 10^{-20} it is safe to remove all table entries below 10^{-100}, for example. On the other hand, the sum cannot be greater than say, $100/a$, otherwise the probability function will exceed 1 by many orders of magnitude. In practice, if a is chosen to be 10, a table of size 40 is sufficient because it covers around 40 orders of magnitude in the density function; for example, from 10^{-30} to 10^{10}.

Note that the first exponential term in the above formula ($B_{ij}^{(1)}$) should not be factored out of the summation. This is to ensure that the correct cancellations occur during the calculation of each product.

5.4 Conclusion

Below is the list of steps necessary for the calculation of the estimation function (without the numerical optimisation discussed above):

- Compute the matrix **S**
- Find the main eigenvector (this defines the rotation matrix **U**)
- Find both eigenvalues S_1 and S_2 using **U** or directly from step 2 (correct S_2 if necessary)
- Find the two bandwidth values h_1 and h_2 using the one-dimensional formula
- Compute matrix $\mathbf{H}_2$
- Find $\mathbf{H}_2^{-1}$ and $\det(\mathbf{H}_2)$ (this is possible because $\det(\mathbf{S}) \neq 0$)
- Build the estimation function $f(x,y)$ using the bivariate normal distribution.

The final density function is automatically normalised to 1.

This approach has been shown to offer better kernel density estimates than the diagonal bandwidth matrix for a number of cases found in the data. As there have not yet been any cases observed that will be less optimally predicted by the non-diagonal bandwidth matrix outlined in this section (when compared to predictions using the diagonal bandwidth matrix), the approach has been adopted for all RSTT guided weapon ground impact data. Our analysis is based on observed cases in the available data, but it is not exhaustive and highlights that the prediction of non-diagonal bandwidth matrices is a challenging and potentially fruitful area of research.

6. Conclusion and areas for further research

In this report we have investigated techniques for analysing simulation data of missile impact distributions.

While we have successfully addressed the fundamental issue of defining techniques of potential value in developing range safety templates, it is clear from the analysis that there are many factors to be considered in generating an accurate statistical model.

In particular, there can be no "blind" fully automated process for producing a statistical estimate that is appropriate for any and every dataset that may arise. It is necessary to ensure that appropriate statistical inputs have been used, and that the results obtained are consistent both with the data and with what might reasonably be expected in reality. Therefore, any procedure for generating a kernel density estimate must be reviewed by a panel of experts, including both statisticians and weapons experts.

Kernel Density Estimation appears to provide a good basis for deriving range safety templates or WDAs from discrete simulated ground impact points.

6.1 Outcomes

The R&D undertaken by CDCIN and DSTO has:

1. Qualitatively described the features of impact distribution data that may affect subsequent statistical modelling.
2. Defined a technique, specifically, the use of kernel density estimation, for providing a statistical model of a specific missile impact data set which estimates the probability density function of the impact distribution. The solution proposed here is purely data analytic and as such does not allow for the incorporation of any substantive knowledge.
3. Defined a technique for combining kernel density estimates corresponding to different failure modes within a single operational scenario.
4. Defined a technique for incorporating information on missile Maximum Energy Boundaries into the analysis so as to refine the impact zone probability density function.
5. Defined a technique for using the probability density function together with population density information to obtain estimated injury rates for a given scenario.
6. Defined a technique for using the probability density function together with range boundary information to obtain an estimate for a missile leaving a given range.
7. Defined a technique for using the probability density function to determine a conservative, convex safety exclusion zone with given probability of the missile leaving the zone.
8. Defined an approximate technique for defining a conservative exclusion zone derived from probability density functions of different scenarios.

9. Found that KDE resolutions beyond 16 x 16 and 32 x 32 do not provide significantly more accurate information and hence 16 x 16 or 32 x 32 resolutions appear to be suitable for the development of Range Safety Templates.
10. Found that at least 600 observations (impact data points) should be used in generating KDEs for a given scenario.
11. Identified situations in which the Kernel Density Estimation process is not robust, generally when tight clusters of data points occur within the data set. In such cases the bandwidth parameters automatically generated by the process tend to be very small and the KDE generated consequently "erratic". This report has suggested one method of dynamic bandwidth calculation to improve the PDF for clustered or non-normal ground impact distributions.
12. Described a covariant form of the Kernel Density Estimator for two-dimensional data that robustly predicts the ground impact probability function for a number of available data sets.
13. Outlined a numerical approach to ensure computationally accurate and efficient results are obtained when using the kernel density estimate technique with real impact data.

6.2 Further research

There are a number of areas in which further research should be carried out in order to make the analysis more robust to the range of potential operational scenarios. The major areas include:

1. The use of kernel density estimation requires selection of certain so-called "bandwidth" parameters. These parameters are the key to controlling the nature of the overall estimate created. In the analysis provided, well-known rules of thumb have been employed, but have been shown to have some potential limitations which should be addressed. Investigation of the implementation of a more general dynamic or adaptive bandwidth method may yield positive results.
2. The number of simulated data points required to generate "sufficiently good" kernel density estimates should be investigated further in light of changes to bandwidth selection algorithms. Additionally, a greater range of impact distribution scenarios should be investigated using a greater number of simulated points.
3. Without significantly greater examination of typical data related to missile impact distributions we are unable at this time to verify that creating a convex exclusion zone from individual scenario exclusion zones will provide a conservative exclusion zone for a scenario whose input parameters lie between the input parameters of the known scenarios. Such an investigation should be carried out in order to determine the full limitations of the defined procedure.
4. Further development of the Exclusion Zone difference to measure the difference between the convex exclusion zones, rather than the raw exclusion zones, for the purposes of comparing variations in KDE algorithms and data sizes.

5. The application of a non-diagonal bandwidth matrix in the kernel density estimation has been shown to be quite useful in a number of observed ground impact data sets. Further exploration of methods for predicting the bandwidth matrix parameters might identify an approach that ensures the optimal kernel density estimate is obtained in most cases.

7. References

1. *TRC Mathematical Modelling* (2004) Range Safety Tool Statistics. DSTO Edinburgh.

2. Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall

3. Silverman, B. W. (1982) *Kernel Density Estimation using the Fast Fourier Transform*. Applied Statistics 31 93-99

4. Scott, D. W. (1992) *Multivariate Density Estimation*, Wiley

5. Zhang, X., King, M. L. and Hyndman, R. J. (2004) *Bandwidth Selection for Multivariate Kernel Density Estimation Using Mcmc*. Clayton, Australia, Monash University

6. Zhang, X., King, M. L. and Hyndman, R. J. (2006) *A Bayesian approach to bandwidth selection for multivariate kernel density estimation*. Computational Statistics & Data Analysis 50 3009-3031

7. Duong, T. and Hazelton, M. L. (2005) *Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation*. Journal of Multivariate Analysis 93 417-433

8. Duong, T. and Hazelton, M. L. (2005) *Cross-validation Bandwidth Matrices for Multivariate Kernel Density Estimation*. Scandinavian Journal of Statistics 32 485-506

This page intentionally left blank

Appendix A: Data files provided by DSTO for the analysis described in this report

Filename	Size	Description
faultset_all_actuators_zeroed_ggm1-10000_1.csv	1.698 MB	10,000 observations corresponding to failure mode 1 (zero deflection) for the spatial scenario chosen to illustrate the techniques proposed in this report.
faultset_all_actuators_zeroed_ggm1-10000_2.csv	1.658 MB	10,000 observations corresponding to failure mode 1 (zero deflection) for a particular spatial scenario.
faultset_all_actuators_zeroed_ggm1-10000_3.csv	1.704 MB	10,000 observations corresponding to failure mode 1 (zero deflection) for a particular spatial scenario.
faultset_all_actuators_zeroed_ggm1-10000_4.csv	1.679 MB	10,000 observations corresponding to failure mode 1 (zero deflection) for a particular spatial scenario.
faultset_all_actuators_zeroed_ggm1-10000_5.csv	1.709 MB	10,000 observations corresponding to failure mode 1 (zero deflection) for a particular spatial scenario.
faultset_all_actuators_zeroed_ggm1-20000_6.csv	3.396 MB	Exactly the same as faultset_all_actuators_zeroed_ggm1-10000_1.csv with a further 10,000 observations added at the beginning (total 20,000 observations).
faultset_single_actuator_freeze_ggm1-20000_3.csv	3.507 MB	20,000 observations corresponding to a different failure mode (single actuator freeze) for a particular spatial scenario.
nofault_ggm_000_1-20000.csv	3.427 MB	20,000 observations corresponding to failure mode 0 (no failure) for the same spatial scenario as faultset_single_actuator_freeze_ggm1-20000_3.csv and faultset_all_actuators_zeroed_ggm1-10000_1.csv.
faultset_all_actuators_zeroed_ggm_007.1-50000.csv	10.038 MB	10,000 for each of five different spatial scenarios which are all identical except for the initial x-distance between the launcher and target.
faultset_all_actuators_zeroed_ggm_008.1-5000.csv	0.99 MB	A further 1000 observations at each of the five scenarios in faultset_all_actuators_zeroed_ggm_007.1-50000.csv corresponding to a different interval of the failure time.
faultset_all_actuators_zeroed_ggm_001.1-480000.csv	94.664 MB	10,000 observations at each of 48 different spatial scenarios. These scenarios consist of all possible combinations of six different initial positions and eight different initial directions of the target.
faultset_all_actuators_zeroed_ggm_009.1-50000.csv	9.522 MB	50,000 observations at a single spatial scenario. This scenario is also symmetric to the scenario in faultset_all_actuators_zeroed_ggm1-10000_1.csv.
faultset_all_actuators_zeroed_ggm_000_1-10000.csv	1.66 MB	Exactly the same as faultset_all_actuators_zeroed_ggm1-10000_1.csv.

This page intentionally left blank

Appendix B: Typical scatter plots and kernel density estimates for samples of various sizes

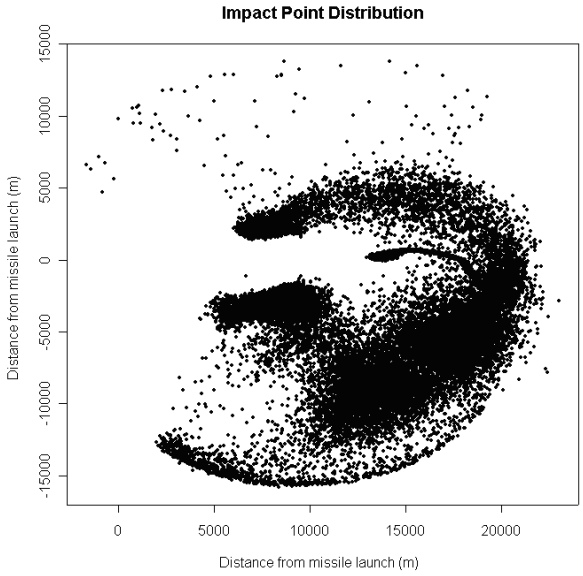

Figure 49: Scatterplot for the dataset of 50,000 observations from the scenario used to investigate the number of points required

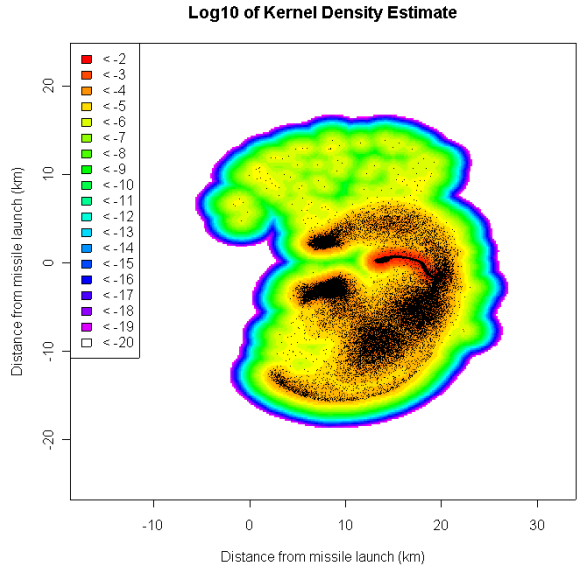

Figure 50: Kernel density estimate for the dataset of 50,000 observations from the scenario used to investigate the number of points required

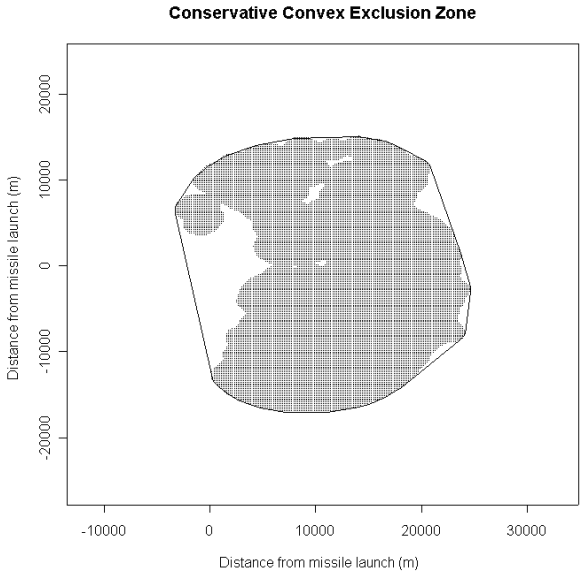

Figure 51: 10^{-6} exclusion zone for the dataset of 50,000 observations from the scenario used to investigate the number of points required

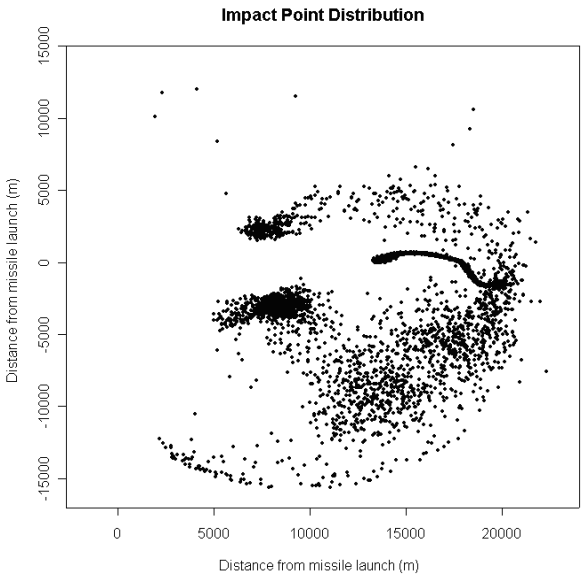

Figure 52: Typical scatterplot for a sample of size 5000 from the scenario used to investigate the number of points required

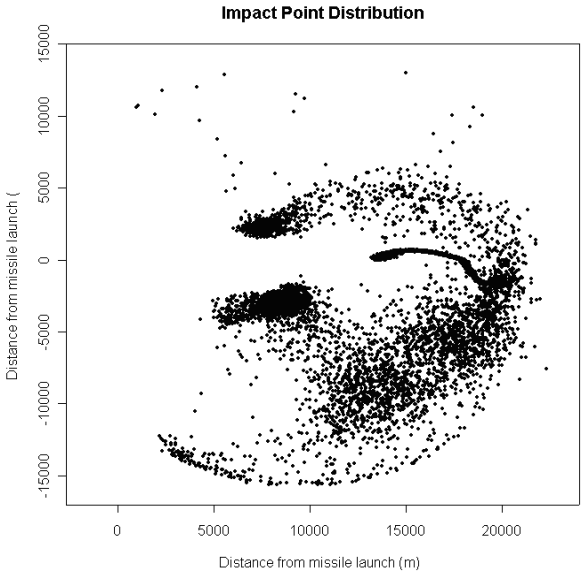

Figure 53: Typical scatterplot for a sample of size 10,000 from the scenario used to investigate the number of points required

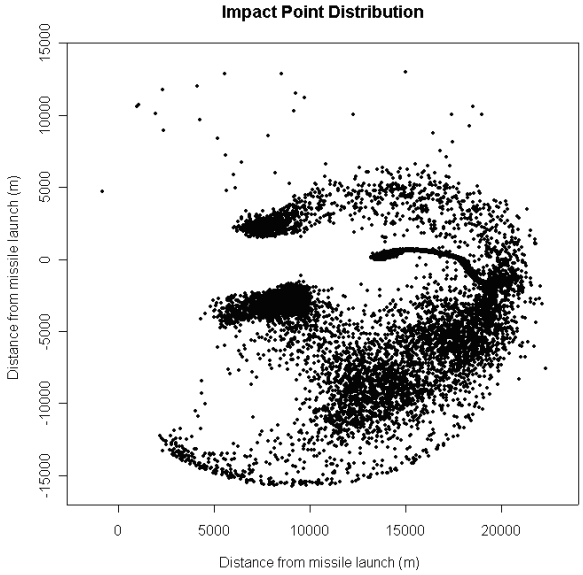

Figure 54: Typical scatterplot for a sample of size 15,000 from the scenario used to investigate the number of points required

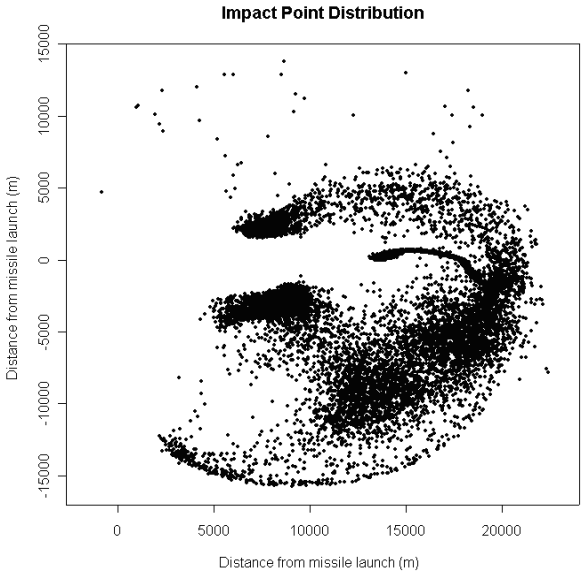

Figure 55: Typical scatterplot for a sample of size 20,000 from the scenario used to investigate the number of points required

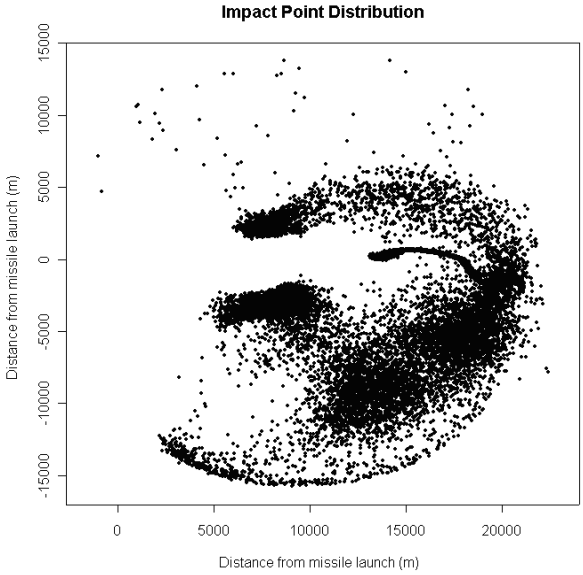

Figure 56: Typical scatterplot for a sample of size 25,000 from the scenario used to investigate the number of points required

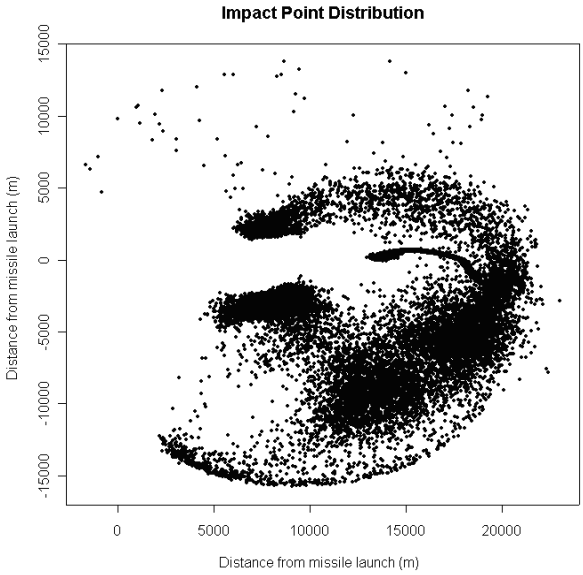

Figure 57: Typical scatterplot for a sample of size 30,000 from the scenario used to investigate the number of points required

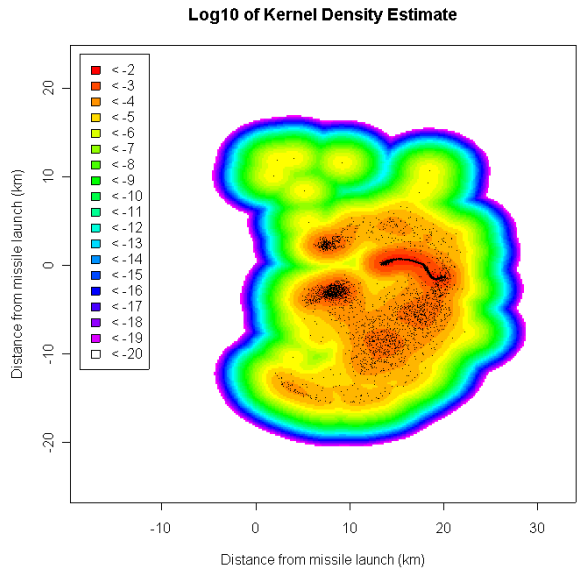

Figure 58: Typical kernel density estimate for a sample of size 5000 from the scenario used to investigate the number of points required

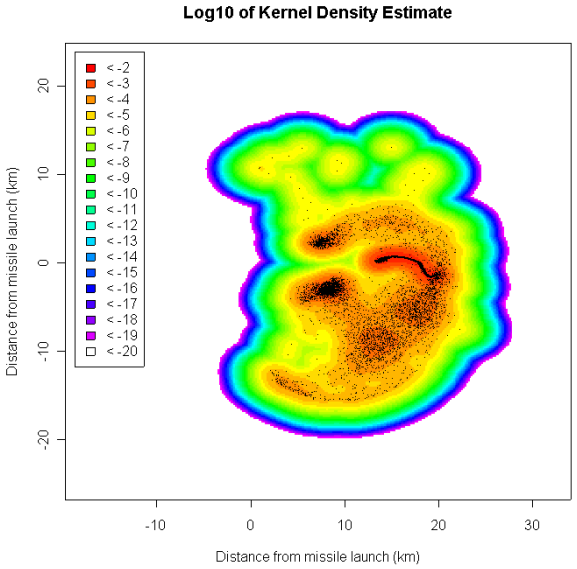

Figure 59: Typical kernel density estimate for a sample of size 10,000 from the scenario used to investigate the number of points required

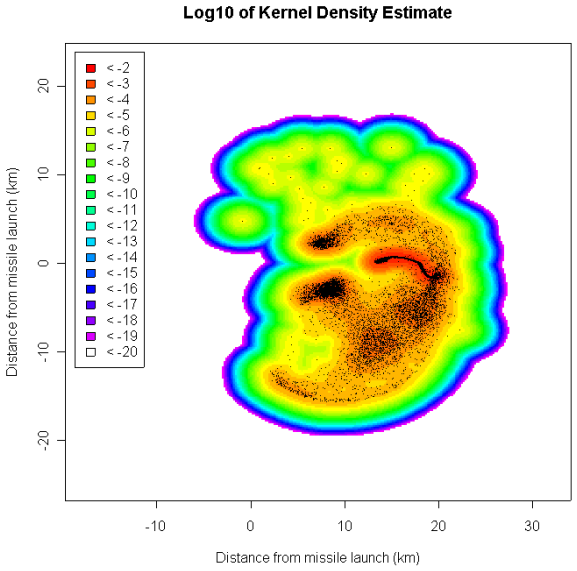

Figure 60: Typical kernel density estimate for a sample of size 15,000 from the scenario used to investigate the number of points required

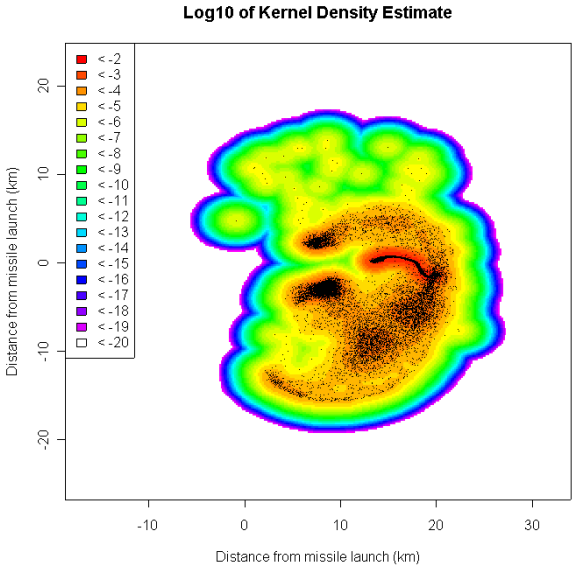

Figure 61: Typical kernel density estimate for a sample of size 20,000 from the scenario used to investigate the number of points required

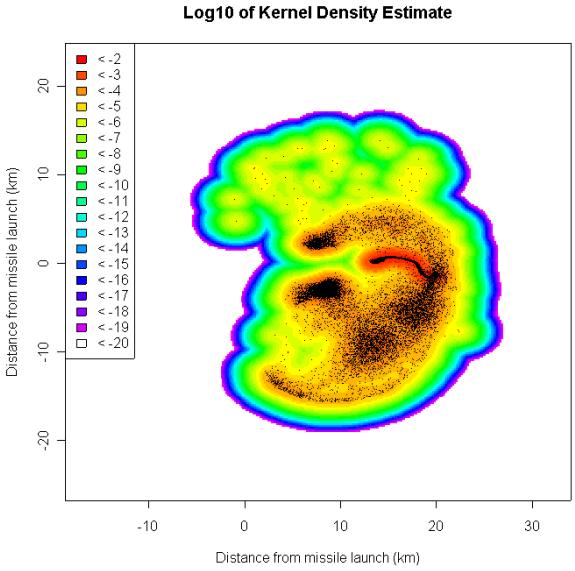

Figure 62: Typical kernel density estimate for a sample of size 25,000 from the scenario used to investigate the number of points required

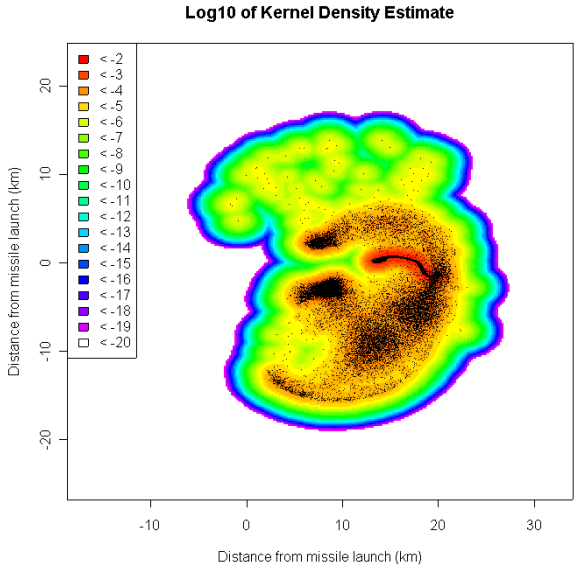

Figure 63: Typical kernel density estimate for a sample of size 30,000 from the scenario used to investigate the number of points required

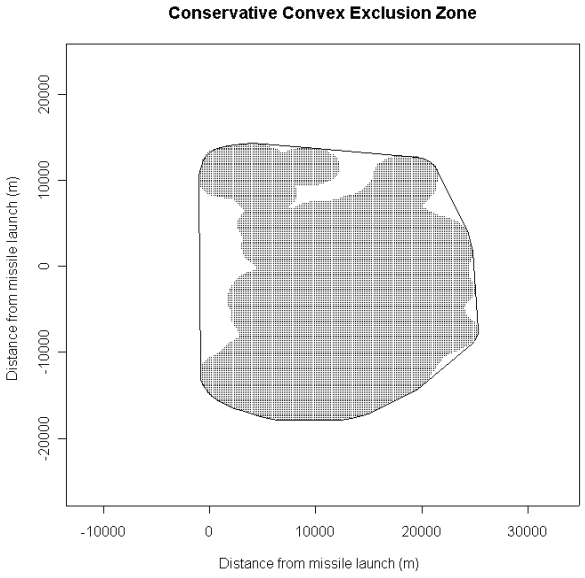

Figure 64: 10^{-6} exclusion zone based on typical kernel density estimate for a sample of size 5000 from the scenario used to investigate the number of points required

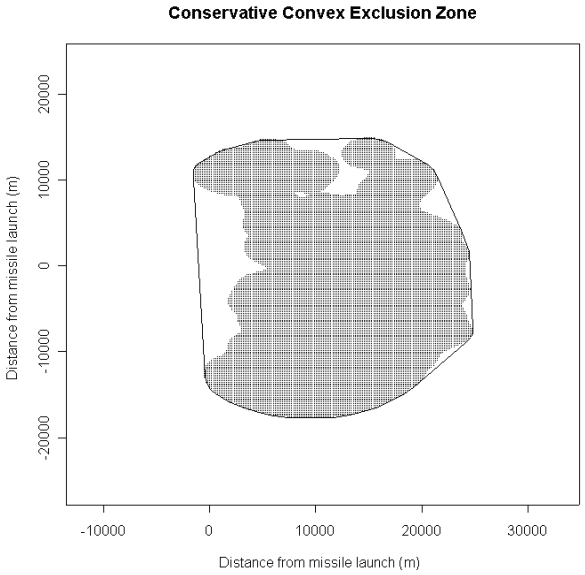

Figure 65: 10^{-6} exclusion zone based on typical kernel density estimate for a sample of size 10,000 from the scenario used to investigate the number of points required

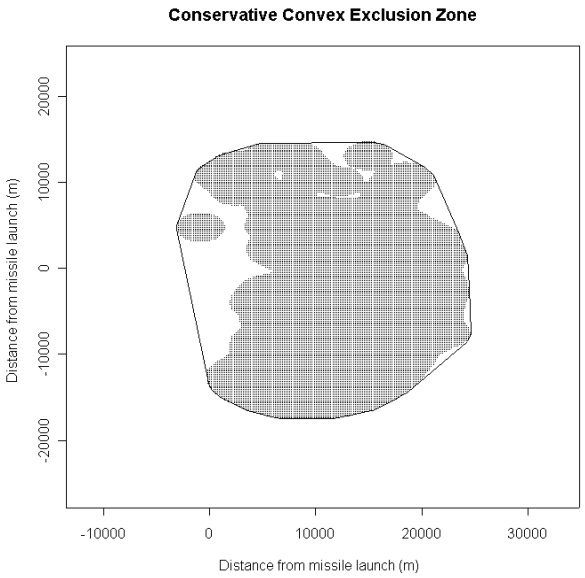

Figure 66: 10^{-6} exclusion zone based on typical kernel density estimate for a sample of size 15,000 from the scenario used to investigate the number of points required

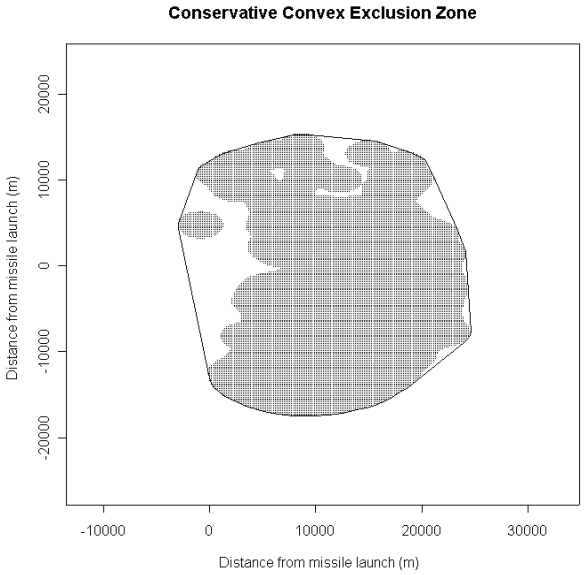

Figure 67: 10^{-6} exclusion zone based on typical kernel density estimate for a sample of size 20,000 from the scenario used to investigate the number of points required

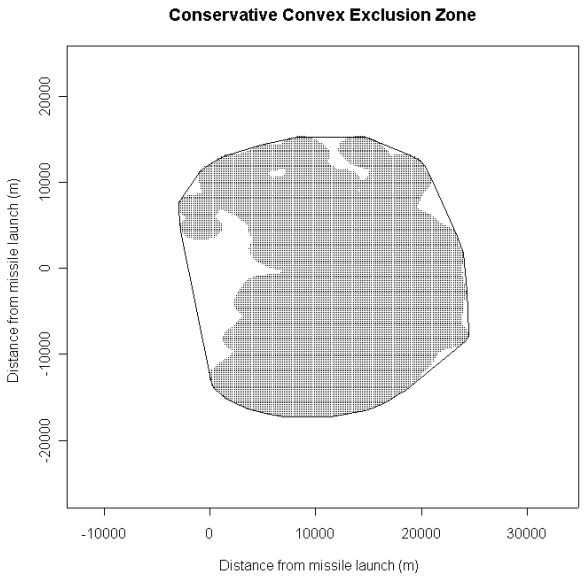

Figure 68: 10^{-6} exclusion zone based on typical kernel density estimate for a sample of size 25,000 from the scenario used to investigate the number of points required

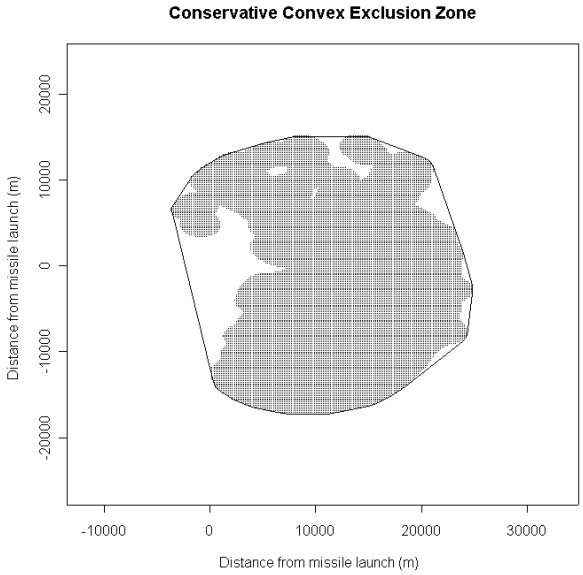

Figure 69: 10^{-6} exclusion zone based on typical kernel density estimate for a sample of size 30,000 from the scenario used to investigate the number of points required

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION
DOCUMENT CONTROL DATA

1. PRIVACY MARKING / CAVEAT (OF DOCUMENT)

2. TITLE

Range Safety Application of Kernel Density Estimation

3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION)

Document	(U)
Title	(U)
Abstract	(U)

4. AUTHOR(S)

Gary Glonek, Timothy Staniford, Michael Rumsewicz,
Oleg Mazonka, Jeremy McMahon, Duncan Fletcher
and Michael Jokic

5. CORPORATE AUTHOR

DSTO Defence Science and Technology Organisation
PO Box 1500
Edinburgh South Australia 5111 Australia

6a. DSTO NUMBER	6b. AR NUMBER	6c. TYPE OF REPORT	7. DOCUMENT DATE
DSTO-TR-2292	AR-014-543	Technical Report	January 2010

8. FILE NUMBER	9. TASK NUMBER	10. TASK SPONSOR	11. NO. OF PAGES	12. NO. OF REFERENCES
2009 / 1069621	AIR 07 / 060	CDRAOSG	71	8

13. URL on the World Wide Web

http: / / www.dsto.defence.gov.au / corporate / reports / DSTO-TR-2292.pdf

14. RELEASE AUTHORITY

Chief, Weapons Systems Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS Yes

18. DSTO RESEARCH LIBRARY THESAURUS http: / / web-vic.dsto.defence.gov.au / workareas / library / resources / dsto_thesaurus.shtml

Research, Methodology, Modelling, Probabilistic modelling

19. ABSTRACT

This report describes the kernel density estimation technique and its application to range safety applications. The kernel density estimation technique is shown to be suitable for developing probabilistic risk assessments from ground impact data generated for guided weapon systems via Monte Carlo simulations. An advantage of this technique is that it can be used to predict the probability density function for minimal simulated ground impacts with apparently random distribution. Several techniques have been proposed to ameliorate the identified limitations of the kernel density estimation technique, including a covariant form for two-dimensional data. Analysis of the available simulated guided weapon ground impact data has identified that around six hundred impact points are sufficient for generating a probability distribution.